

**НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»**

ФАКУЛЬТЕТ ІНФОРМАТИКИ ТА ОБЧИСЛЮВАЛЬНОЇ ТЕХНІКИ

Кафедра автоматизованих систем обробки інформації і управління

«На правах рукопису»

«До захисту допущено»

В.о. завідувача кафедри

УДК 004.75

О.А.Павлов
(підпис) (ініціали, прізвище)

Магістерська дисертація

зі спеціальності 121 «Інженерія програмного забезпечення»

на тему: «Інтелектуальна система класифікації осіб на основі даних соціальних мереж»

Виконав:

студент VI курсу, групи *ІП-82мв*

Кахерський О.І.

(прізвище, ім'я, по батькові)

(підпис)

**Науковий
керівник**

доц., доц., к.ф.-м.н. Гавриленко О.В.,

(посада, науковий ступінь, вчене звання, прізвище та ініціали)

(підпис)

Консультант

доц., к.т.н., Ліщук К.І.

(посада, науковий ступінь, вчене звання, прізвище та ініціали)

(підпис)

Рецензент

професор, д.ф.-м.н., професор Безверхий О.І.

(посада, науковий ступінь, вчене звання, прізвище та ініціали)

(підпис)

Засвідчую, що у цій магістерській дисертації немає запозичень з праць інших авторів без відповідних посилань.

Студент

(підпис)

Київ – 2019 року

**Національний технічний університет України
«Київський політехнічний інститут
імені Ігоря Сікорського»**

Факультет інформатики та обчислювальної техніки
(повна назва)

Кафедра автоматизованих систем обробки інформації та управління
(повна назва)

Рівень вищої освіти другий (магістерський) за освітньо-професійною програмою

Спеціальність 121 «Інженерія програмного забезпечення»
(код і назва)

ЗАТВЕРДЖУЮ

В.о. завідувача кафедри

 О.А. Павлов
(підпис) (ініціали, прізвище)

«___» _____ 201 р.

ЗАВДАННЯ
на магістерську дисертацію студенту
Кахерському Олегу Ігоровичу
(прізвище, ім'я, по батькові)

1. Тема дисертації Інтелектуальна система класифікації осіб на основі даних соціальних мереж

науковий керівник дисертації к.ф.-м.н., доц. Гавриленко О.В.
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

затверджені наказом по університету від “___” _____ 20__ р. № _____

2. Строк подання студентом дисертації “___” _____ 20__ р.

3. Об'єкт дослідження Процес класифікації осіб на основі даних соціальних мереж

4. Предмет дослідження Методи та підходи до процесу класифікації осіб за зображеннями, отриманими із соціальних мереж

5. Перелік завдань, які потрібно розробити аналіз існуючих автоматичних та напіваавтоматичних методів ідентифікації осіб;

вирішення задачі автоматичної ідентифікації осіб на зображеннях;

вирішення задачі класифікації ідентифікованих осіб;

реалізація програмного застосунку класифікації осіб за необробленими зображеннями, отриманими із соціальних мереж

6. Перелік графічного матеріалу

Функціональна схема процесу класифікації користувача

Функціональна схема передачі управління під час класифікації користувача

Структурна схема сервісу групування користувачів

7. Орієнтовний перелік публікацій *Підхід до розв'язання задач класифікації осіб на основі даних, отриманих із соціальних мереж*

Модифікація алгоритму для вирішення задач ідентифікації облич

8. Консультанти розділів дисертації

| Розділ | Прізвище, ініціали та посада консультанта | Підпис, дата | |
|-----------|---|----------------|------------------|
| | | завдання видав | завдання прийняв |
| Графічний | Ліщук К.І., доц. | | |
| | | | |
| | | | |

9. Дата видачі завдання “ 01 ” вересня 20 19 р

Календарний план

| № з/п | Назва етапів виконання магістерської дисертації | Строк виконання етапів магістерської дисертації | Примітка |
|-------|---|---|----------|
| 1 | Аналіз існуючих автоматичних та напіваавтоматичних методів ідентифікації осіб | 20.06.2019 | |
| 2 | Вирішення задачі автоматичної ідентифікації осіб на зображеннях | 30.09.2019 | |
| 3 | Вирішення задачі класифікації ідентифікованих осіб на зображеннях | 30.09.2019 | |
| 4 | Реалізація програмного забезпечення класифікації осіб за необробленими зображеннями, отриманими із соціальних мереж | 30.10.2019 | |
| 5 | Проведення експериментальних досліджень розробленого алгоритму | 30.11.2019 | |
| 6 | Оформлення документації | 05.12.2019 | |
| 7 | Подання роботи на попередній захист | 05.12.2019 | |
| 8 | Подання роботи на основний захист | 18.12.2019 | |

Студент

(підпис)

(ініціали, прізвище)

Науковий керівник

(підпис)

(ініціали, прізвище)

РЕФЕРАТ

Магістерська дисертація: 84с., 12рис., 18табл., 3 додатка, 14 джерел.

Актуальність теми: Обравши будь-якого середньостатистичного користувача соціальних мереж, можна отримати базові відомості його життя, лише переглянувши його сторінку, але якщо почати більш ґрунтовно аналізувати його публічні дії в мережі, то можна зробити більш цінні висновки. Створення системи, яка дозволить виконувати моніторинг та аналіз хоча б окремих аспектів діяльності користувачів в соціальних мережах може слугувати додатковим джерелом для прийняття рішень бізнесу чи урядові, а також суспільним організаціям.

Мета дослідження: Основна мета роботи полягає в розробці системи, яка виконуватиме класифікацію осіб на основі зображень, отриманих із соціальних мереж.

Для реалізації поставленої мети були сформульовані наступні завдання:

- виконати аналіз існуючих автоматичних та напівавтоматичних методів ідентифікації осіб;
- вирішити задачу автоматичної ідентифікації осіб на зображеннях;
- вирішити задачі напівавтоматичної класифікації ідентифікованих осіб на зображеннях;
- реалізувати програмний застосунок класифікації осіб за зображеннями, отриманими із соціальних мереж.

Об'єкт дослідження: Процес класифікації осіб на основі даних соціальних мереж.

Предмет дослідження: Методи та підходи до процесу класифікації осіб за зображеннями, отриманими із соціальних мереж.

Методи дослідження: При проведенні досліджень використовувалися методи знаходження ключових точок на зображеннях, критеріїв подібності для порівняння різних зображень, теорії графів та алгоритмів знаходження спільнот в мережах для групування зображень, а також напівавтоматичні методи для класифікації користувачів за результатами групування зображень.

Наукова новизна: Найбільш суттєвими науковими результатами магістерської дисертації є:

- запропоновано алгоритм автоматичної ідентифікації осіб на зображеннях;
- запропонована модифікація критерію порівняння зображень на основі локальних детекторів ознак.

Практичне значення отриманих результатів визначається тим, що була спроектована та реалізована система, що дозволяє виконати класифікацію осіб за даними фотографій із соціальних мереж

Зв'язок роботи з науковими програмами, планами, темами. Робота виконувалась в рамках науково-дослідних робіт кафедри автоматизованих систем обробки інформації та управління факультету обчислювальної техніки та інформатики КПІ імені Ігоря Сікорського № 0117U000914 «Математичні моделі та технології в системах підтримки прийняття рішень».

Апробація: Основні положення роботи доповідались і обговорювались на Всеукраїнській науково-практичній конференції молодих вчених та студентів «Інформаційні системи та технології управління» (ІСТУ-2019).

Публікації. Матеріали роботи опубліковані в тезах Всеукраїнської науково-практичної конференції молодих вчених та студентів «Інформаційні системи та технології управління» (ІСТУ-2019) та тезах доповідей Дванадцятої міжнародної науково-технічної конференції «Проблеми інформатизації».

Ключові слова: Розпізнавання обличчя, ідентифікація обличчя, локальні детектори ознак, методи знаходження клік, методи знаходження спільнот, напіваавтоматична класифікація.

ABSTRACT

Master dissertation: 84 pp., 12 fig., 18 tab., 3 app., 14 sources.

Topicality: Choosing any average social network user, you can get basic information about his life just by looking at his page, but if you start looking more closely at his public activities online, you can draw more valuable conclusions. Creating a system that will allow monitoring and analysis of at least some aspects of user activity on social networks can serve as an additional source for decision-making by business or government, as well as by public organizations.

The purpose of the research: The main purpose of the study is to develop a system for individuals classification based on images fetched from social networks.

To achieve the purpose the next tasks were formulated:

- to revise existing automatic and semi-automatic methods for individuals recognition;
- to solve the task of automatic individuals recognition;
- to solve the task of semi-automatic classification of individuals recognized in images;
- to provide a software for individuals classification based on images fetched from social networks.

The object of study: The process of individuals classification based on social networks data.

The subject of the research: methods and approaches of the process of individuals classification based on social networks data

Methods of the research: The research used methods of finding key points on images, similarity criteria for comparing different images, graph theory and algorithms for finding communities in networks to group similar images, semi-automatic methods of individuals classification based on results of grouping images.

Scientific novelty: The most significant scientific results of the dissertations:

- an algorithm for automatic individuals recognition in images has been proposed;
- a criteria of image comparison based on local feature detectors has been modified.

The practical significance of the results is a system designed and implemented to allow individuals based on images fetched from social networks.

Relationship of work with scientific programs, plans, themes: The research was carried out at the Department of Computer-Aided Management And Data Processing Systems of the National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute» within the theme «Mathematical models and technologies in decision support systems» State registration number 0117U000914.

Approbation: The main points of the work were reported and discussed on All-Ukrainian Scientific and Practical Conference of Young Scientists and Students "Information Systems and Technologies of Management" (ISTU-2019).

Publications: The materials of the thesis are published in the theses of the All-Ukrainian Scientific and Practical Conference of Young Scientists and Students "Information Systems and Technologies of Management" (ISTU-2019) and of the Twelfth International Scientific and Technical Conference “Problems of Informatization”.

Keywords: Face detection, face recognition, local feature detectors, clique finding methods, community detection methods, semi-automatic classification.

ЗМІСТ

| | |
|---|----|
| ЗМІСТ | 8 |
| ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ..... | 11 |
| ВСТУП..... | 13 |
| РОЗДІЛ 1 ОПИС ІСНУЮЧИХ МЕТОДІВ ТА ТЕХНОЛОГІЙ | 16 |
| 1.1 Методи попередньої обробки зображення | 16 |
| 1.1.1 Перетворення зображення з RGB формату в напівтоновий..... | 16 |
| 1.1.2 Зменшення шумів | 17 |
| 1.1.3 Модифікація гистограми зображення | 18 |
| 1.2 Задача ідентифікації об'єктів на зображенні | 21 |
| 1.3 Опис локальних дескрипторів ознак при вирішенні задачі порівняння .. | 23 |
| 1.3.1 Опис методу SIFT | 24 |
| 1.3.2 Обчислення відстані в задачах класифікації..... | 29 |
| 1.3.3 Коефіцієнти подібності | 29 |
| 1.4 Вирішення задачі знаходження спільнот | 30 |
| 1.4.1 Загальні відомості..... | 30 |
| 1.4.2 Алгоритм Брона-Кербоша знаходження клік | 31 |
| 1.4.3 Модифікація алгоритму Брона-Кербоша з вибором опорної вершини ... | 32 |
| 1.4.4 Модифікація алгоритму Брона-Кербоша з впорядкуванням вершин..... | 33 |
| 1.4.5 Опис алгоритму Label Propagation..... | 33 |
| Висновки до розділу..... | 34 |

| | |
|--|----|
| РОЗДІЛ 2 МАТЕМАТИЧНЕ ЗАБЕЗПЕЧЕННЯ..... | 37 |
| 2.1 Вирішення задачі автоматичної ідентифікації осіб на зображеннях..... | 37 |
| 2.1.2 Опис процесу ідентифікації | 37 |
| 2.1.3 Модифікація критерію подібності Кернексі | 38 |
| 2.1.4 Модифікація алгоритму Брона-Кербоша для ідентифікації осіб | 41 |
| 2.2 Вирішення задачі напівавтоматичної класифікації осіб | 44 |
| 2.2.1 Постановка задачі | 44 |
| 2.2.2 Опис запропонованого рішення..... | 44 |
| Висновки до розділу..... | 45 |
| РОЗДІЛ 3 РЕАЛІЗАЦІЯ СИСТЕМИ..... | 46 |
| 3.1 Постановка задачі..... | 46 |
| 3.2 Контекст роботи системи | 46 |
| 3.3 Опис складових частин системи | 48 |
| 3.4 Опис Супервізора | 52 |
| 3.5 Опис Конвеєра підготовки даних | 53 |
| 3.6 Опис Конвеєра обчислення степені схожості користувачів | 54 |
| 3.7 Опис Конвеєра з групування користувачів..... | 55 |
| 3.8 Опис Конвеєра класифікації користувачів..... | 58 |
| 3.9. Опис Менеджера даних..... | 58 |
| Висновки до розділу..... | 62 |
| РОЗДІЛ 4 ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ..... | 63 |
| 4.1 Загальні відомості..... | 63 |

| | |
|--|----|
| 4.2 Дослідження ефективності модифікованого критерію подібності при порівнянні облич з базою даних облич | 63 |
| 4.3 Дослідження ефективності методу автоматичної ідентифікації осіб | 64 |
| 4.4 Дослідження ефективності роботи методу класифікації..... | 64 |
| 4.4 Відомості про швидкодію роботи системи | 65 |
| Висновки до розділу..... | 65 |
| РОЗДІЛ 5 РОЗРОБЛЕННЯ СТАРТАП-ПРОЕКТУ | 66 |
| 5.1 Опис ідеї проекту..... | 66 |
| 5.2 Технологічний аудит проекту | 66 |
| 5.3 Аналіз ринкових можливостей запуску стартап-проекту | 69 |
| 5.4 Розроблення ринкової стратегії проекту | 76 |
| Висновки до розділу..... | 78 |
| ЗАГАЛЬНІ ВИСНОВКИ..... | 79 |
| СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ..... | 80 |
| ДОДАТОК А ФУНКЦІОНАЛЬНА СХЕМА ПРОЦЕСУ КЛАСИФІКАЦІЇ КОРИСТУВАЧА..... | 82 |
| ДОДАТОК Б ФУНКЦІОНАЛЬНА СХЕМА ПРОЦЕСУ ПЕРЕДАЧІ УПРАВЛІННЯ ПІД ЧАС КЛАСИФІКАЦІЇ КОРИСТУВАЧА..... | 83 |
| ДОДАТОК В СТРУКТУРНА СХЕМА СЕРВІСУ ГРУПУВАННЯ КОРИСТУВАЧІВ..... | 84 |

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

RGB – red, green, blue. Формат кольорового зображення, кожен колір якого складається з суми кольорів червоного, зеленого та синього каналів.

Perceptual luminance-preserving (luminosity) method – колориметричний метод конвертації RGB зображення в напівтоновий формат, виконується зважене сумування RGB каналів.

Blurring – процес розмиття зображення, тобто зменшення його чіткості видимості його контурів.

Mean filter – фільтр середнього арифметичного, використовується для розмиття зображення.

Weighted mean filter – фільтр зваженого середнього арифметичного, використовується для розмиття зображення.

Histogram stretching – процес розтягнення гістограми зображення, тим самим збільшуючи його контрастність.

LoG – Лапласіан гаусіанів (англ. Laplacian of Gaussian), детектор кутів.

DoG – Різниця гаусіанів (англ. Difference of Gaussian), детектор кутів.

HoG – Гістограма орієнтованих градієнтів (англ. Histogram of oriented gradients), локальний дескриптор ключових точок.

LBP – Локальні бінарні шаблони (англ. local binary patterns), локальний дескриптор ключових точок.

SIFT – Масштабонезалежне перетворення ознак (англ. scale-invariant feature detector). Локальний дескриптор ключових точок.

PCA-SIFT – principal component analysis – SIFT, алгоритм використання методу головних компонент разом з SIFT.

SURF – speeded up robust feature, локальний дескриптор ключових точок.

SDK – набір із засобів розробки, утиліт і документації, що дозволяє створювати прикладні програми за визначеною технологією або для певної платформи.

API – прикладний програмний інтерфейс (англ. Application Programming Interface).

JVM – віртуальна машина Java (англ. Java Virtual Machine).

ВСТУП

Актуальність теми.

Якщо обрати будь-якого середньостатистичного користувача соціальних мереж, то можна отримати базові відомості його життя, лише переглянувши його сторінку: ПІБ, дата народження та місце проживання, стать та національність, сімейний стан та освіту, взаємозв'язки з іншими користувачами даної мережі. Дана інформація сама по собі не несе особливої цінності, але якщо почати більш ґрунтовно аналізувати його публічні дії в мережі, то можна зробити більш цінні висновки:

- на новій доданій фотографії можна ідентифікувати інших користувачів мережі та виконати ранжирування його друзів по степені близькості з даним користувачем, проаналізувавши частоту появи сумісних фотографій;
- по “лайкам” та коментарям до фотографій та постів можна дізнатись, з ким користувач взаємодіє більше в мережі зі своїх друзів та підписників, а з ким взагалі не підтримує публічний зв'язок;
- проаналізувавши активні підписки користувача та його діяльність в спільнотах, можна зробити відповідні висновки щодо його вподобань чи поглядів на ту чи іншу соціально значиму подію.

Однак, опрацювати такі колосальні об'єми даних, які примножуються з величезною швидкістю, а також просто відслідкувати оновлення чи іншу активність користувача з подальшим аналізом в ручному режимі дуже складно, а інколи – просто неможливо. У зв'язку з цим, створення системи, яка дозволить виконувати моніторинг та аналіз хоча б окремих аспектів діяльності користувачів в соціальних мережах може слугувати додатковим джерелом для прийняття рішень бізнесу чи урядові, а також суспільним організаціям.

Зв'язок роботи з науковими програмами, планами, темами.

Дисертаційна робота виконувалась в рамках науково-дослідних робіт кафедри автоматизованих систем обробки інформації та управління факультету обчислювальної техніки та інформатики КПП імені Ігоря Сікорського № 0117U000914 «Математичні моделі та технології в системах підтримки прийняття рішень».

Мета і задачі дослідження.

Основна мета роботи полягає в розробці системи, яка виконуватиме класифікацію осіб на основі зображень, отриманих із соціальних мереж.

Для досягнення поставленої мети необхідно виконати наступні задачі:

- виконати аналіз існуючих автоматичних та напівавтоматичних методів ідентифікації осіб;
- вирішити задачу автоматичної ідентифікації осіб на зображеннях;
- вирішення задачі напівавтоматичної класифікації ідентифікованих осіб на зображеннях;
- реалізувати програмний застосунок класифікації осіб за необробленими зображеннями, отриманими із соціальних мереж.

Об'єкт дослідження.

Процес класифікації осіб на основі даних соціальних мереж.

Предмет дослідження.

Методи та підходи до процесу класифікації осіб за зображеннями, отриманими із соціальних мереж.

Методи дослідження.

При проведенні досліджень використовувалися методи знаходження ключових точок на зображеннях, критеріїв подібності для порівняння різних зображень, теорії графів та знаходження спільнот в мережах для групування зображень, а також напівавтоматичні методи для класифікації користувачів за результатами групування зображень. Для експериментального дослідження використовувалися технології обробки надвеликих

масивів даних для розподіленої підготовки масивів зображень та їх порівняння, об'єктно-орієнтованого програмування для реалізації компонентів системи, генерації структури соціальних мереж і композиції зображень із датасетів ORL та MegaFace.

Наукова новизна одержаних результатів полягає в наступному:

- виконана модифікація критерію порівняння зображень на основі детекторів локальних ознак;
- розроблений алгоритм ідентифікації осіб на зображеннях в автоматичному режимі.

Практичне значення одержаних результатів.

Розроблена система, що здатна виконувати класифікацію осіб на основі необроблених даних із соціальних мереж.

Публікації.

Матеріали роботи опубліковані в тезах третьої всеукраїнської науково-практичної конференції молодих вчених та студентів «Інформаційні системи та технології управління» (ІСТУ-2018) – м. Київ.: НТУУ «КПІ ім. Ігоря Сікорського» та тезах доповідей Дванадцятої міжнародної науково-технічної конференції «Проблеми інформатизації».

Ключові слова.

Розпізнавання обличчя, ідентифікація обличчя, локальні детектори ознак, методи знаходження клік, методи знаходження спільнот, напіваавтоматична класифікація.

РОЗДІЛ 1 ОПИС ІСНУЮЧИХ МЕТОДІВ ТА ТЕХНОЛОГІЙ

1.1 Методи попередньої обробки зображення

1.1.1 Перетворення зображення з RGB формату в напівтоновий

З метою спрощення подальшої обробки кольорового зображення його зазвичай переводять у напівтоновий формат. Це пояснюється тим, що кожен піксель кольорового зображення представляється 3 числами i , таким чином, для обробки зображення необхідно виконувати операції над 3 матрицями. У зв'язку з цим, зображення в RGB форматі конвертується в напівтоновий формат і наступна обробка виконується лише над єдиною матрицею (через те, що кожен піксель зображення відтінком сірого, тобто одним числом).

Існує два способи конвертації RGB в напівтоновий формат:

- метод середнього арифметичного;
- колориметричний (англ. perceptual luminance-preserving, luminosity method).

Метод середнього арифметичного заключається в знаходженні середнього арифметичного інтенсивності червоного, зеленого та синього кольорів пікселю. Результат такої конвертації зображено на рисунку 1.1.



Рисунок 1.1 – Результати конвертації кольорового зображення в напівтоновий формат методом середнього арифметичного

Як видно з рисунку 1.1 основним недоліком даного методу є те, що кожен колір має однаковий внесок у формування зображення. В той же час, кожен з трьох кольорів має різну довжину хвилі і через це внесок у формування зображення в них є різним.

Для врахування недоліку, описаного вище, використовується колориметричний метод конвертування кольорового зображення в напівтоновий. Через те, що червоний колір має більшу довжину хвилі, зелений – меншу і вносить заспокійливий ефект для очей, то внесок зеленого кольору повинен бути збільшений, червоного – зменшений, а синього – між двома останніми. У зв'язку з цим формула перетворення набуває наступного вигляду [1]:

$$\text{Grayscale} = (0.3 * R) + (0.59 * G) + (0.11 * B) \quad (1.1)$$

Таким чином, внесок в результуючу інтенсивність відтінку сірого зеленого кольору збільшена до 59%, синього зменшена до 11% і червоного зменшена до 30%. Результат такої конвертації зображено на рисунку 1.2.



Рисунок 1.2 – Результат конвертації кольорового зображення в напівтоновий формат колориметричним методом

1.1.2 Зменшення шумів

Одним з основних етапів підготовки зображення для подальшої обробки, пов'язаної з пошуком контурів об'єктів є зменшення шумів зображення. Одним із засобів для

зменшення шумів є розмиття зображення (англ. blurring). Наприклад, зображення обличчя виглядає чітко, коли ми здатні розпізнати його елементи (наприклад, очі, ніс, вуха), тому концепція розмиття дозволяє залишити лише більш чіткі основні контури, зменшивши кількість побічних [2].

Розмиття можна досягти за допомогою використання фільтрів, найбільш поширеними з яких є фільтр середнього арифметичного (англ. mean filter), фільтр зваженого середнього арифметичного (англ. weighted average filter) та фільтр Гауса. Фільтри використовуються в операції згортки.

1.1.3 Модифікація гістограми зображення

Гістограмою цифрового зображення з рівнями інтенсивності в діапазоні $[0, L-1]$ (де L – загальна кількість рівнів інтенсивності) називається дискретна функція $h(r_k) = n_k$, де r_k є k -й рівень інтенсивності, а n – число пікселів на зображенні, що мають інтенсивність r_k . Загальною практикою є нормування гістограми шляхом ділення кожного її значення на загальну кількість пікселів на зображенні, що позначається добутком MN , де M і N – розміри зображення. Таким чином, розміри нормованої гістограми будуть $p(r_k) = n_k/MN$ для $k = 0, 1, \dots, L-1$, а $p(r_k)$ – імовірність появи пікселя зі значенням інтенсивності r_k . [3, с. 162]

Гістограми є основою для багатьох методів просторової обробки. Видозміна гістограми може бути використана для покращення якості зображення.

Однією з поширеною модифікацією гістограми є її зміщення. Такий ефект досягається додаванням константи до інтенсивності всіх пікселів на зображенні. Це дозволяє змінювати яскравість зображення. Додавши 50 до інтенсивності зображення, було отримано результат, зображений на рисунку 1.3, а віднявши 80 – на рисунку 1.4.

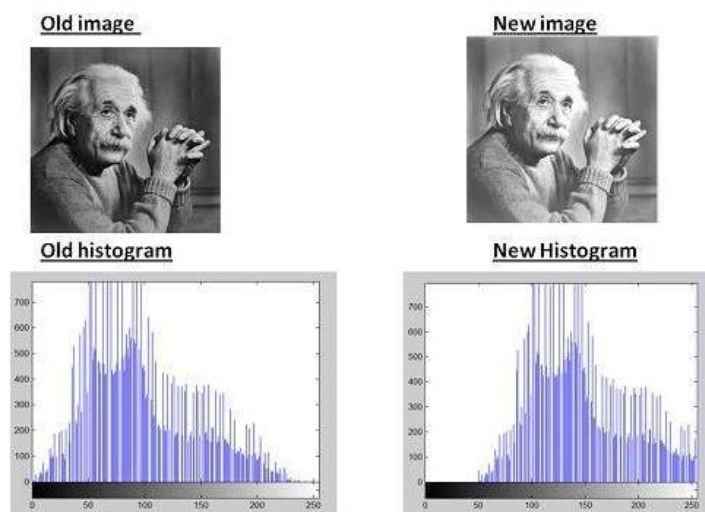


Рисунок 1.3 – Результат збільшення яскравості зображення шляхом додавання 50 до інтенсивності всіх пікселів. Зліва розміщено вхідне зображення та його гістограма, зправа – модифіковане зображення з гістограмою

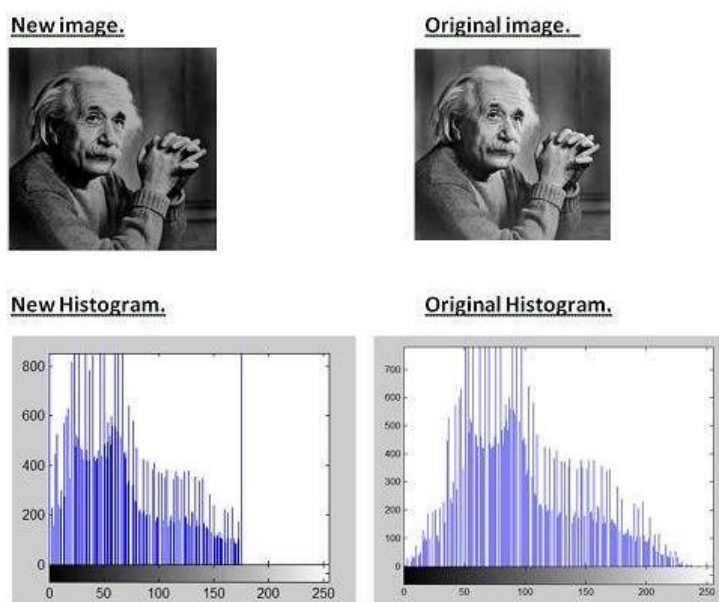


Рисунок 1.4 – Результат зменшення яскравості зображення шляхом віднімання 80 від інтенсивності всіх пікселів. Зліва розміщено вхідне зображення та його гістограма, зправа – модифіковане зображення з гістограмою

Іншим видом модифікації зображення є зміна його контрастності. Контрастністю зображення називається різниця між його максимальною та мінімальною інтенсивностями. Такий ефект досягається шляхом процесу, який називається розтягнення гістограми (англ. Histogram stretching).

Розтягнення гістограми виконується за допомогою використання формули перетворення 1.2 на кожному пікселі зображення:

$$g(x, y) = \frac{f(x, y) - f_{min}}{f_{max} - f_{min}} * 2^k, \quad (1.2)$$

де $f(x, y)$ – поточне значення інтенсивності пікселя;

k – кількість рівнів сірого.

Ще одним способом покращення якості зображення за допомогою модифікації його гістограми є стабілізація гістограми (частотна корекція, англ. histogram equalization). Він дозволяє покращити контрастність зображення. Слід зауважити, що для цього не завжди необхідно збільшувати контрастність.

Алгоритм виконання стабілізації гістограми:

- обчислити функцію імовірності по всіх пікселях зображення;
- виконати кластеризацію на L класів, де L – кількість відтінків сірого;
- обчислити кумулятивну функцію імовірності;
- для кожного кластеру обчислити наступне значення кластеру за формулою $CDF * (L-1)$, де CDF – значення функції імовірності в даному кластері;
- для кожного кластеру виконати переміщення частот в наступний обчислений кластер.

Результат стабілізації гістограми для різномірних зображень зображено на рисунку 1.5.

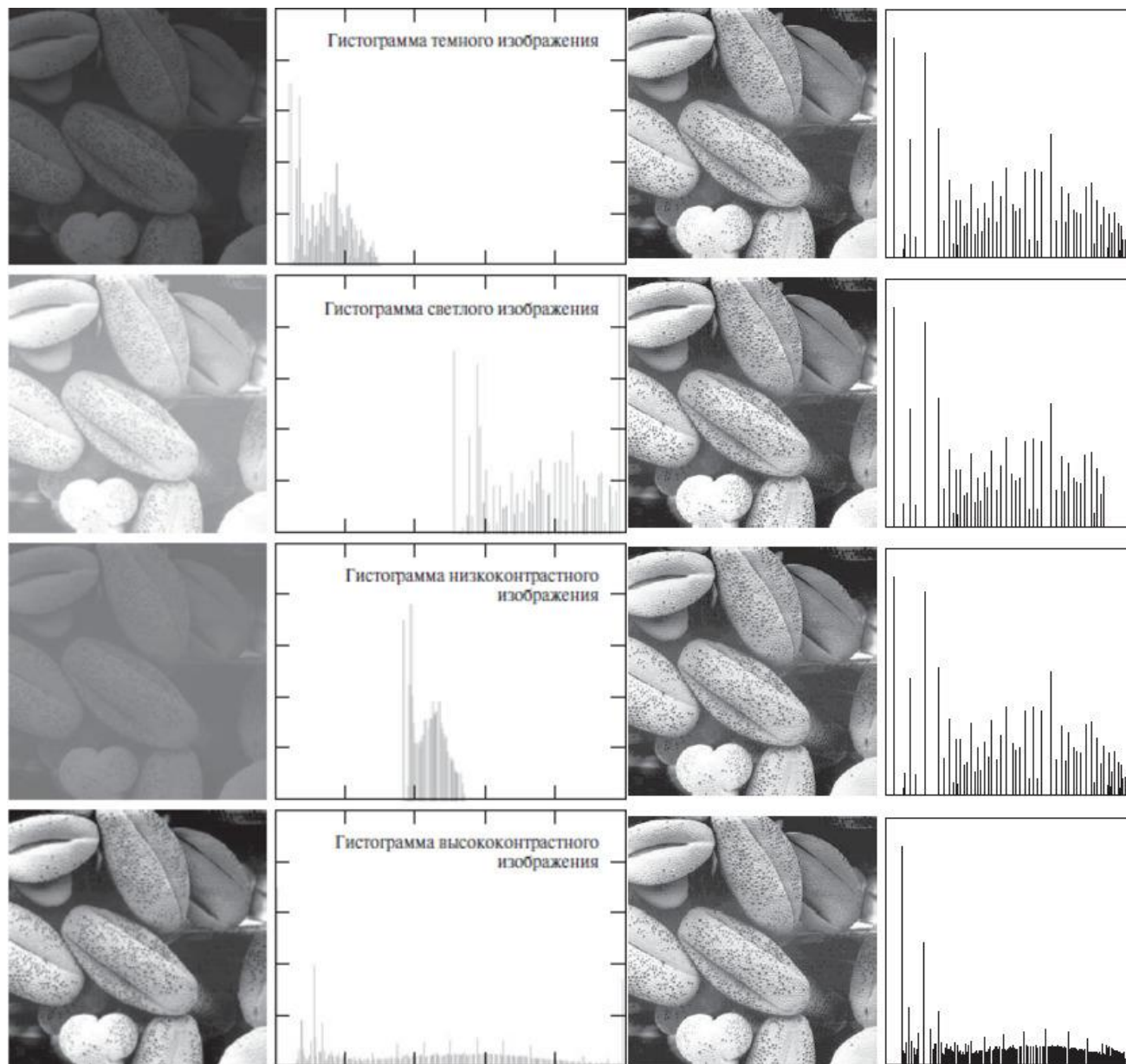


Рисунок 1.5 – Результати стабілізації гистограми темного, світлого, низькоконтрастного та висококонтрастного зображень [3]

1.2 Задача ідентифікації об'єктів на зображенні

В деяких задачах з комп'ютерного зору необхідно виконувати пошук співпадаючих об'єктів на декількох кадрах або зображеннях. Цього можна досягти виконавши відповідний пошук співпадаючих точок. Наприклад, знайшовши декілька співпадаючих точок між двома кадрами стереозображення можна з легкістю співставити майже всі інші

точки на зображенні (точно так же, як це робить наш мозок). Ця задача може бути вирішена за допомогою так званих ознак (англ. feature)[4]. Також слід зауважити, що відображення зображення на глобальну ознаку не є гарною практикою, замість виконується пошук декількох локальних ознак, адже зміна, наприклад, несуттєвої частини зображення впливатиме на результат порівняння.

Загальні вимоги до локальної ознаки:

- повторювана: однакові точки незалежно розпізнаються на кожному зображенні;
- інваріантність до поворотів та масштабування;
- інваріантність до афінних трансформацій;
- інваріантність до наявності шумів та розмиття;
- локальність: надійність до наявності перешкод та зміни освітлення;
- відмінність: регіон повинен мати “цікаву” структуру;
- кількість: повинно бути достатньо точок для представлення зображення;
- часова ефективність.

Найлегшим способом пошуку об’єкту на зображенні є виконання співставлення з шаблонами. Основним обмеженням у практичному використанні є те, що необхідно мати шаблон для співставлення на момент роботи алгоритму, що є не завжди можливим.

Більш складним способом є виконання наступної послідовності кроків:

- знаходження ключових на зображенні (кутів, граней і т.д.);
- відображення даних точок та регіону навколо них у локальну ознаку за допомогою дескрипторів;
- виконання співставлення локальних ознак.

Деякі популярні детектори кутів:

- оператор Гарріса;
- лапласіан гаусіанів (англ. Laplacian of Gaussian (LOG));
- різниця гаусіанів (англ. Difference of Gaussian (DOG));

- детектор Лапласіанів Гарріса (англ. Harris Laplacian detector).

Приклад роботи детектору Гессе (англ. Hessian affine region detector) та оператору Гарріса зображено на рисунку 1.6.

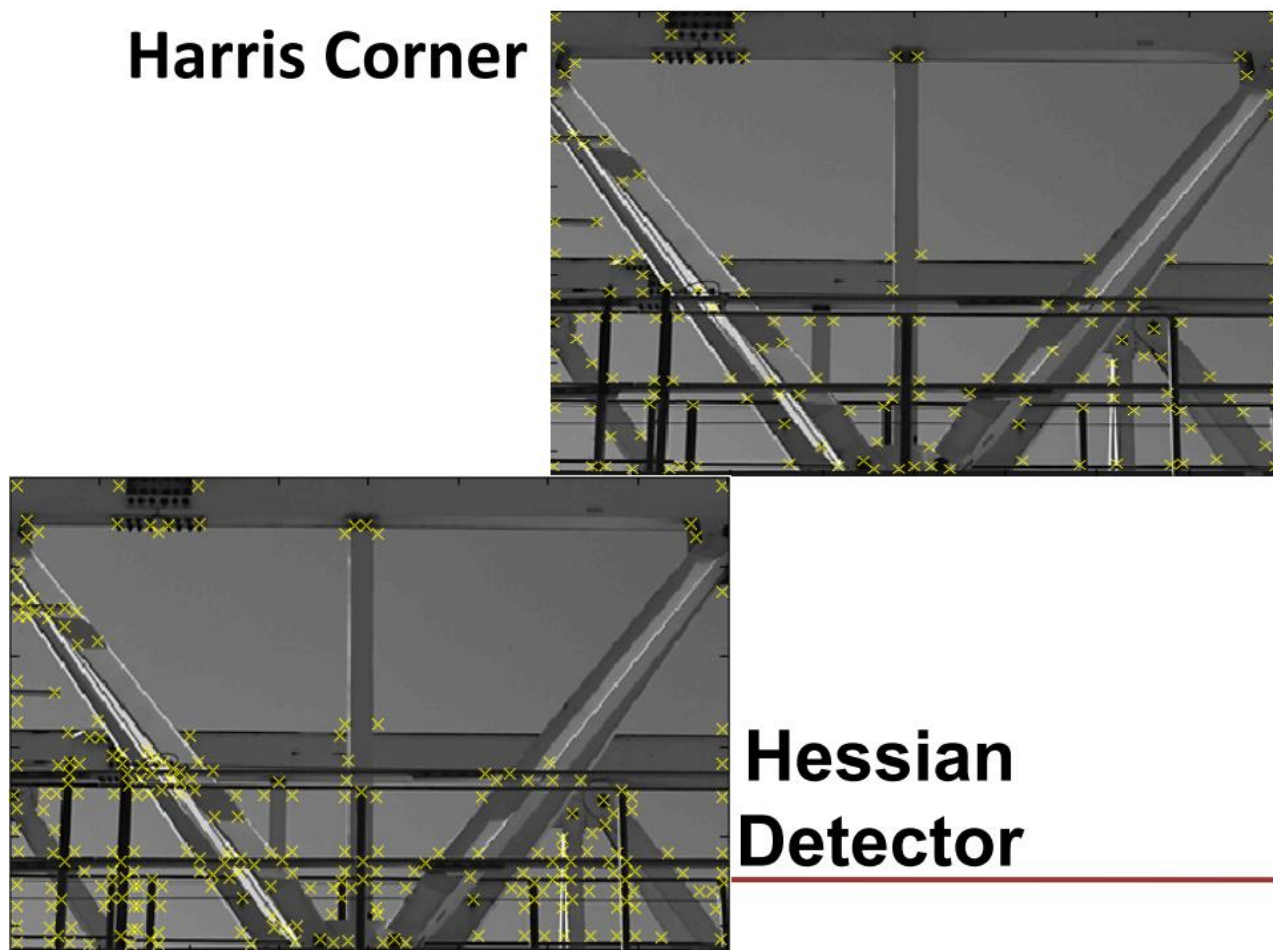


Рисунок 1.6 – Результат роботи оператору Гарріса та детектору Гессе

1.3 Опис локальних дескрипторів ознак при вирішенні задачі порівняння

Як було зазначено в п.1.2 застосування локальних дескрипторів значно підвищує надійність розпізнавання об'єкту на зображенні.

Найбільш відомі локальні дескриптори:

- гістограма орієнтованих градієнтів (англ. histogram of oriented gradients, HOG);

- локальні бінарні шаблони (англ. local binary descriptors, LBP);
- масштабнонезалежне перетворення ознак (англ. scale-invariant feature detector, SIFT);
- PCA-SIFT (метод головних компонент з SIFT, англ. principal component analysis);
- SURF (speeded up robust feature).

1.3.1 Опис методу SIFT

SIFT - алгоритм із області комп'ютерного зору, який виявляє і описує локальні ознаки зображення. Він використовується, коли необхідно виконати розпізнавання об'єкту на зображеннях, які за своєю природою можуть мати різний масштаб, освітлення та кути нахилу.

Алгоритм складається з наступних операцій:

- побудова простору масштабів;
- LoG апроксимація;
- пошук ключових точок;
- фільтрація знайдених точок;
- визначення орієнтації ключових точок;
- генерація SIFT ознак.

Для формування простору масштабів необхідно початкове зображення зменшити в 2 рази і повторити цю операцію 4 рази. Таким чином, отримуємо 5 зображень, кожне наступне з яких в 2 рази менше, ніж попереднє.

Наступним кроком є застосування Гаусового розмиття на кожному з зображень 5 разів, виконати операцію згортки кожного зображення з оператором Гауса за формулою 1.3:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y), \quad (1.3)$$

де L – “розмите” зображення;

G – оператор Гауса;

I – вхідне зображення;

x, y – координати місцезнаходження;

σ – рівень розмиття, чим більше значення, тим більше розмиття зображення;

$*$ - оператор згортки;

Оператор Гауса обчислюється за формулою 1.4:

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (1.4)$$

Слід зазначити, що степінь розмиття кожного наступного зображення відрізняється від попереднього на $k\sigma$, де k – будь-яка константа.

Таким чином, внаслідок застосування останніх двох операцій вхідне зображення відображається на чотири групи по п'ять зображень. Кожна група називається октавою.

Приклад формування простору масштабів зображено на рисунку 1.7.

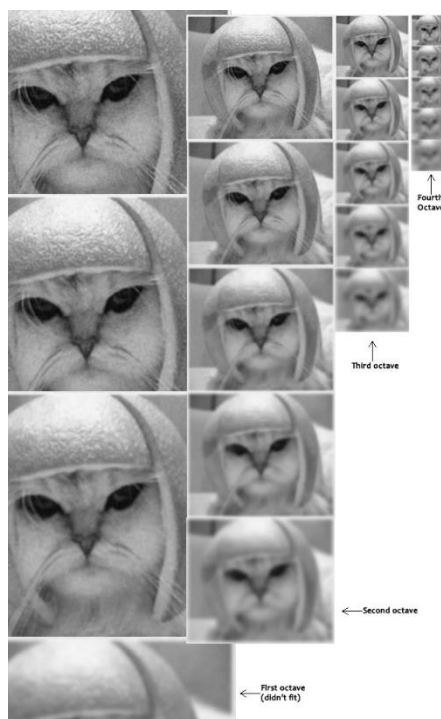


Рисунок 1.7 – Приклад формування простору масштабів для довільного зображення

Кількість октав повинно бути обрано в залежності від розмірів зображення, але автор у своїй праці зазначає, що п'яти октав є достатньою умовою для більшості задач.

Для ефективного пошуку стабільних ключових точок було запропоновано використання різниці гаусіанів, яка може бути обчислена як різниця двох сусідніх зображень в межах октави (згідно з формулою 1.5):

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) = L(x, y, k\sigma) - L(x, y, \sigma). \quad (1.5)$$

Існує декілька причин використання даної функції:

- L повинна бути обчислена в будь-якому випадку для опису масштабонезалежної ознаки і D може бути обчислена як різниця зображень;
- різниця гаусіанів є наближенням до нормалізованої функції лапласіану гаусіанів і тому зменшуються затрати на обчислення, адже використовується просте віднімання.

Приклад обчислення різниці гаусіанів для вхідного зображення зображено на рисунку 1.8.

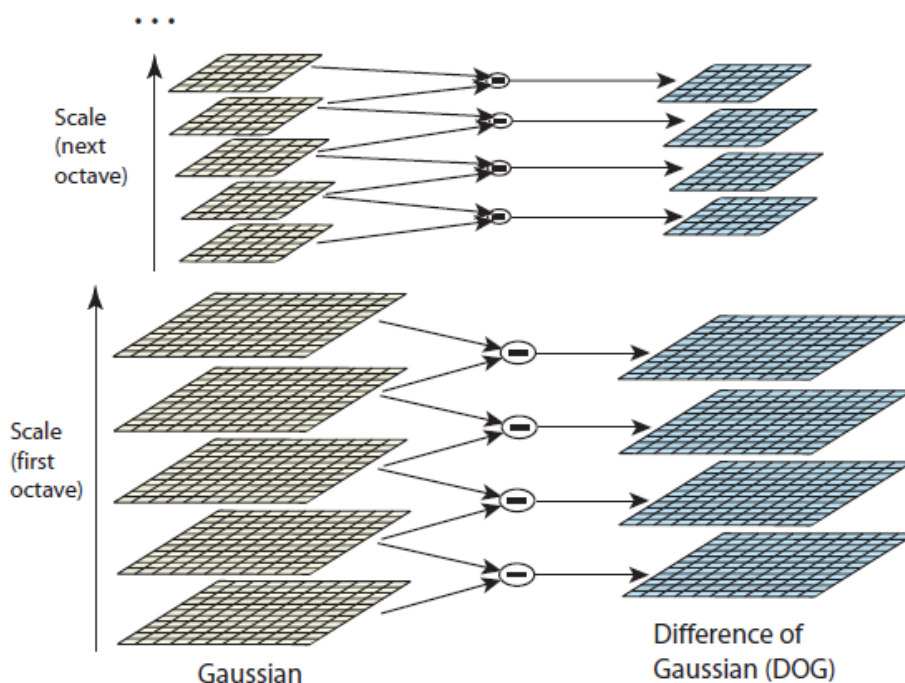


Рисунок 1.8 – Обчислення різниці гаусіанів для сусідніх зображень кожної октави

Далі для пошуку ключових точок необхідно знайти точки екстремуму. Для цього необхідно порівняти кожен піксель з його сусідніми на поточному зображенні, а також на верхньому та нижньому. В цілому необхідно виконати 26 порівнянь. Поточна точка є ключовою, якщо вона більша чи менша за всі сусідні. На практиці достатньо виконати декілька порівнянь, щоб відкинути точку як ключову. Процес зображено на рисунку 1.9.

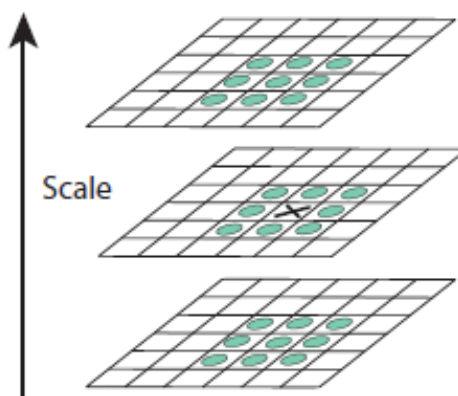


Рисунок 1.9 – Процес знаходження точки екстремуму

Після цього необхідно обчислити значення субпікселя за допомогою розкладу в ряд Тейлора зображення навколо ключової точки та прирівнення виразу до нуля. Після обчислення буде знайдено положення субпікселя, що робить алгоритм більш стабільним та збільшує шанси на знаходження співпадінь.

Наступним кроком є фільтрація низькоконтрасних ключових точок, а також плоских регіонів та граней, таким чином, залишаючи лише кутові точки.

Далі необхідно визначити напрямок ключової точки. Для цього необхідно обчислити градієнт та величини навколо ключової точки. Потім знаходиться найчастіші значення напрямків в регіоні. Розмір регіону навколо ключової точки залежить прямопропорційно до масштабу. Напрямок та значення градієнту обчислюються за формулами 1.6-1.7:

$$m(x, y) = \sqrt{(L(x + 1, y) - L(x - 1, y))^2 + (L(x, y + 1) - L(x, y - 1))^2}, \quad (1.6)$$

$$\theta(x, y) = tg^{-1} \left(\frac{L(x, y + 1) - L(x, y - 1)}{L(x + 1, y) - L(x - 1, y)} \right) \quad (1.7)$$

Останнім кроком є визначення вектору ознак з ключової точки. Для цього необхідно виконати наступні кроки:

- визначити область 16x16 навколо ключової точки;
- розбити визначену область на 16 областей 4x4;
- обчислити абсолютне значення градієнту та його орієнтацію в кожній з областей;
- виконати кластеризацію отриманих результатів на 8 значень та створити гістограму. Значення, додане до кластеру гістограми, залежить від абсолютного значення градієнту і відстані від ключової точки. Для цього використовується фільтр Гауса;
- виконати нормалізацію значень вектору. [5]

Схематично цей процес зображено на рисунку 1.10.

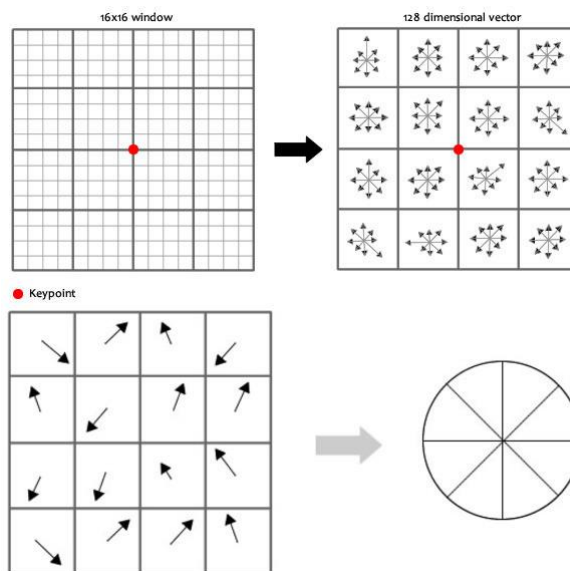


Рисунок 1.10 – Процес відображення ключової точки на 128-розмірний вектор

1.3.2 Обчислення відстані в задачах класифікації

Поняття відстані є одним із ключових в задачах класифікації. Від вибору метрики можуть бути змінені результати класифікації, а отже – і правильність роботи моделі.

Нижче наведені найбільш поширені метрики:

- евклідова відстань, яка визначається як довжина різниці векторів, що представляють об'єкти, що порівнюються;
- манхеттенська відстань, яка визначається як сума абсолютних поєдинатних різниць двох векторів;
- відстань Геммінга, яка визначається як кількість неспівпадаючих елементів двох векторів.

1.3.3 Коефіцієнти подібності

Коефіцієнт подібності (також міра подібності, індекс подібності) - безрозмірний показник подібності порівнюваних об'єктів. Також відомий під назвами «міра асоціації», «міра подібності» і ін.

Застосовується в біології для кількісного визначення ступеня подібності біологічних об'єктах (ділянках, районів, окремих фітоценозів, зооценозів і т. д.). Також застосовуються в географії, соціології, розпізнаванні образів, пошукових системах, порівняльній лінгвістики, біоінформатики, хемоінформатика, при порівнянні рядків і ін.

У більш широкому сенсі говорять про заходи близькості до яких відносяться: заходи різноманітності, заходи концентрації (однорідності), заходи включення, міри схожості, заходи відмінності (в тому числі відстані), заходи сумісності подій, заходи несумісності подій, заходи взаємозалежності. Теорія заходів близькості знаходиться в стадії становлення і тому існує безліч різних уявлень про формалізацію відносин близькості.

Більшість коефіцієнтів нормовані і знаходяться в діапазоні від 0 (схожість відсутній) до 1 (повна схожість). Подібність і відмінність взаємодоповнюють один одного (математично це можна висловити так: схожість = 1 - різниця).

Коефіцієнти подібності можна умовно розділити на три групи в залежності від того, яка кількість об'єктів розглядається:

- унарні - розглядається один об'єкт. У цю групу входять заходи різноманітності і заходи концентрації;
- бінарні - розглядається два об'єкти. Це найбільш відома група коефіцієнтів;
- n-арні (багатомісні) - розглядається n об'єктів. Ця група найменш відома.

1.4 Вирішення задачі знаходження спільнот

1.4.1 Загальні відомості

В теорії складних мереж вважається, що мережа має структуру спільноти, якщо вузли можуть бути легко зібрані в групи (можливо такі, що мають спільні вузли) таким чином, що всередині груп вузли будуть тісно пов'язаними. У випадку спільнот, що не перетинаються, із цього слідує, що мережа ділиться на групи вузлів з тісними зв'язками всередині спільноти та розрідженими – між групами. В той же час спільноти, що мають спільні вузли, також є допустимими. Більш загальне визначення базується на тому, що пара вузлів з більшою вірогідністю матиме зв'язок у випадку, якщо вони належать до одної спільноти.

Існуючі методи знаходження спільнот можна класифікувати наступним чином:

- на основі окремих вузлів:
 - 1) використання критеріїв подібності;
 - 2) обчислення степені вершин;
 - 3) врахування досяжності вершин;
- на основі груп:
 - 1) врахування модульності;
 - 2) на основі збалансованості;
 - 3) ієрархічні;
 - 4) врахування щільності.

Найвідоміші методи для знаходження спільнот:

- алгоритм поширення міток (англ. Label Propagation algorithm);
- знаходження оптимальної модульності;
- провідний власний вектор (англ. Leading Eigenvector);
- багаторівневий алгоритм (англ. Multi-Level);
- швидкий жадібний алгоритм (англ. Fast-Greedy);
- алгоритм випадкового блукання (англ. Walktrap);
- InfoMAP;
- алгоритм Брона-Кербоша та його модифікації.

1.4.2 Алгоритм Брона-Кербоша знаходження клік

Алгоритм Брона-Кербоша – це алгоритм для знаходження максимальних клік в неорієнтованому графові. Таким чином, він перераховує всі підмножини вершин з двома властивостями, що кожна пара вершин в одній із перелічених підмножин – з'єднана ребром, і жодна перелічена підмножина не може додавати до неї будь-які додаткові вершини, зберігаючи її повну сполучуваність. Хоча інші алгоритми для вирішення проблеми клік мають час роботи, який, теоретично, є кращим для вхідних даних, які мають декілька максимальних незалежних множин, даний алгоритм та його вдосконалення часто повідомляються як більш ефективні на практиці, ніж альтернативи [10].

Основна форма алгоритму Брона-Кербоша - це рекурсивний алгоритм пошуку з поверненням, який шукає всі максимальні кліки в заданому графіку G . Загалом, з урахуванням трьох непересічних наборів вершин R , P і X , він знаходить максимальні кліки, які включають усі вершин у R , частину вершин у P , жодну з вершин у X . У кожному виклику алгоритму P і X - це множини, що не пересікаються, об'єднання яких складається з тих вершин, які утворюють кліки при додаванні до P . Інакше кажучи, $P \cup X$ - це

сукупність вершин, які приєднуються до кожного елемента R . Коли P і X порожні, більше немає елементів, які можна додати до R , тому R є максимальною клікою і алгоритм виводить R .

Рекурсія ініціюється встановленням R і X пустими множинами, а P - сукупністю вершин графу. У межах кожного рекурсивного виклику алгоритм по черзі розглядає вершини P ; якщо таких вершин немає, він або повертає R як максимальну кліку (якщо X порожня), або повертається на крок назад. Для кожної вершини v , обраної з P , він робить рекурсивний виклик, в якому v додається R і в якому P і X складаються з сусідів $N(v)$ з v , який знаходить і повідомляє про всі розширення кліки R , які містять v . Потім він переміщує v від P до X , щоб виключити його з розгляду в майбутніх кліках, і продовжує роботу з наступною вершиною в P [11].

1.4.3 Модифікація алгоритму Брона-Кербоша з вибором опорної вершини

Основна форма алгоритму, описана вище, є неефективною у випадку графів з великою кількістю не максимальних клік: він виконує рекурсивний виклик для кожної кліки, максимальної чи ні. Для заощадження часу і швидшого проходження по гілкам пошуку, які не містять максимальних клік, Брон і Кербош запропонували варіант алгоритму, який включає в себе вибір опорної вершини u з множини P (або, як пізніше було досліджено в [12], з $P \cup X$). Будь-яка максимальна кліка повинна включати або u , або одного з її сусідів, інакше кліку можна доповнити, додавши до неї u . Отже, лише u та його не сусідні вершини потрібно перевірити як вибір для вершини v , яка додається до R у кожному рекурсивному виклику алгоритму.

Вибір опорної вершини дозволяє зменшити кількість рекурсивних викликів алгоритму, що дозволяє значно зменшити час роботи алгоритму в порівнянні з базовим варіантом алгоритму [13].

Згідно з дослідженнями в [12] асимптотична складність модифікації – $O(3^{\frac{n}{3}})$.

1.4.4 Модифікація алгоритму Брона-Кербоша з впорядкуванням вершин

Альтернативний метод вдосконалення основної форми алгоритму Брона-Кербоша передбачає вибір вершин з упорядкованої множини P на першому рівні рекурсивного виклику замість вибору опорної вершини з метою мінімізації розмірів множин P вершин-кандидатів у межах кожного рекурсивного виклику.

Виродженість графу G – найменше число d , таке, що кожен підграф G має вершини зі ступенями d або менше. Кожен граф має порядок виродження, впорядкування вершин таким чином, що кожна вершина має d або менше сусідів, які приходять пізніше при впорядкуванні; порядок виродження може бути знайдено за лінійний час шляхом багаторазового вибору вершини мінімального ступеня серед вершин, що залишилися. Якщо порядок вершин v , через який проходить алгоритм Брона-Кербоша, - це порядок виродження, то множина P вершин-кандидатів у кожному виклику (сусіди v , які згодом будуть впорядковані) гарантовано матиме розмір не більше d . Множина X виключених вершин буде складатися з усіх попередніх сусідів v і може бути значно більшою, ніж d . У рекурсивних викликах до алгоритму нижче першого рівня рекурсії все ще може бути використана модифікація з вибором опорної вершини [14].

Згідно з дослідженнями в [14] асимптотична складність модифікації - $O\left(dn3^{\frac{d}{3}}\right)$, де n – кількість вершин графу, d – виродженість графу.

1.4.5 Опис алгоритму Label Propagation

Алгоритм поширення міток - це алгоритм часткового навчання, який призначає мітки попередньо не класифікованим даним. На початку алгоритму (як правило, невелика) підмножина точок даних має мітки. Ці мітки поширюються на нерозмічені точки протягом усього алгоритму.

У складних мережах реальні мережі мають структуру спільноти. Поширення міток - це алгоритм для пошуку спільнот. У порівнянні з іншими алгоритмами розповсюдження міток має переваги у часі роботи та кількості апріорної інформації, необхідної про структуру мережі (заздалегідь не потрібно знати жодного параметра). Недоліком є те, що воно не дає унікального рішення, а сукупність багатьох рішень. Також результати роботи алгоритму можуть відрізнятися один від одного при незмінних вхідних даних [15].

Згідно з [15], базовий процес має 5 етапів.

Етап 1. Ініціалізація міток в графові за формулою 1.8:

$$C_x(0) = x \quad (1.8)$$

Етап 2. Встановлення $t = 1$.

Етап 3. Впорядкування вузлів випадковим чином в графові та утворення множини X .

Етап 4. Обчислення значення нової мітки C для кожного $x \in X$ за формулою 1.9:

$$C_x(t) = f(C_{x_{i1}}(t), \dots, C_{x_{im}}(t), C_{x_{i(m+1)}}(t-1), \dots, C_{x_{ik}}(t-1), \quad (1.9)$$

де f повертає мітку, що виникає з найвищою частотою серед сусідів. У випадку наявності декількох міток з найвищою частотою, мітка обирається випадковим чином.

Етап 5. Якщо кожен вузол має мітку, яку має максимальна кількість їх сусідів, то алгоритм завершує роботу. В іншому випадку виконується збільшення значення t і далі виконується етап 3.

Висновки до розділу

Задача розпізнавання образів в розділі комп'ютерного зору є досить типовою, на що вказують багато наукових праць з високою кількістю цитувань. Загалом її вирішення зводиться до підготовки зображення до аналізу за допомогою методів цифрової обробки зображень, таких як модифікування гістограми зображення, зміни його розмірів, вирізання його частин для зменшення часу та стабільності роботи алгоритму, зменшення

шумів за допомогою процедури розмиття. Після цього можна використати один із запропонованих локальних дескрипторів ознак для відображення зображення на масив ключових точок, кожна з яких описана за допомогою дескриптору. Такий підхід дозволяє збільшити стабільність роботи методів розпізнавання образів, адже забезпечується інваріантність до поворотів об'єкту, наявності перешкод та зміни в освітленні.

Головним обмеженням, що накладається на цей вид задач, є те, що розпізнавання виконується на заздалегідь підготовленій базі даних зображень. Таким чином, ця задача добре вирішується методами машинного навчання з учителем в той час, коли за відсутності бази даних якість ідентифікації є достатньо низькою.

Знайшовши можливість звести задачу ідентифікації до задачі знаходження спільнот та проаналізувавши існуючі методи їх знаходження, можна стверджувати, що методи знаходження спільнот не задовольняють вимоги задачі через те, що не накладають обмеження на якість співпадіння зв'язків всередині спільнот, а також повноту цих зв'язків. З іншого боку, алгоритми знаходження максимальних клік, тобто повних підграфів, накладають дуже жорсткі умови на такі спільноти, адже в повному графові зв'язки між усіма вершинами мають відповідати критеріям якості. Таке рішення є не досить гнучким, адже в результаті роботи механізму порівняння можуть виникати аномальні результати або ж дані можуть бути не повні. Іншим недоліком є те, що один елемент може належати різним максимальним клікам, що є недопустимим для вирішення задачі ідентифікації облич.

Задача класифікації вирішується за допомогою методів навчання з учителем або без (через методи кластеризації), але вони на вхід отримують уже підготовлені статичні дані, тобто дані, які, в цілому, не змінюються під час роботи алгоритму. У випадку поставленої цілі одним із критеріїв класифікації є аналіз взаємозв'язків користувача, якого необхідно класифікувати, а також результатів класифікації інших користувачів. Таким чином, задача класифікації в даному випадку залежить від порядку обробки користува-

чів, а також повноти даних для обробки і повною відсутністю тестової вибірки для виконання навчання класифікатора. Ці обмеження та особливості не дозволяють використовувати алгоритми навчання з вчителем, а також найбільш поширені алгоритми класифікації (навчання без учителя), тому необхідно розробити метод або адаптувати існуючий під умови та обмеження поставленого завдання.

РОЗДІЛ 2 МАТЕМАТИЧНЕ ЗАБЕЗПЕЧЕННЯ

2.1 Вирішення задачі автоматичної ідентифікації осіб на зображеннях

2.1.1 Постановка задачі

Дано набір зображень, необхідно автоматично ідентифікувати осіб на них, якщо загальна кількість унікальних осіб – невідома.

2.1.2 Опис процесу ідентифікації

Як було зазначено в першому розділі, задача розпізнавання може бути зведена до послідовного виконання наступних кроків:

- знаходження обличчя особи на зображенні;
- підготовка зображення;
- відображення зображення в уніфікований вигляд;
- порівняння поточного зображення із базою даних зображень;
- знаходження найбільшого значення критерію співпадіння;
- прийняття рішення щодо прийняття співпадіння між зображеннями.

Так як база даних облич відсутня за умовами поставленої задачі, порівняння зображення відбувається з усіма наявними зображеннями в системі, окрім тих, що були знайдені на одному зображенні з поточним. Замість знаходження найбільшого значення критерію співпадіння вводиться порогове значення, на основі якого можна зробити висновки, що знайдено обличчя однієї особи.

У випадку знаходження обличчя на зображенні виконуються наступні кроки підготовки:

- вирізання зображення обличчя;
- відображення в напівтоновому форматі;
- нормалізація розмірів зображення;
- зменшення шумів на зображенні;

- нормалізація гістограми (англ. histogram equalization).

Наступним етапом є відображення зображення в уніфікований вигляд, який буде вхідним параметром для критерію подібності.

Як було зазначено в першому розділі, достатньо популярними є методи знаходження локальних точок на зображенні, створення на їх основі дескрипторів та порівняння їх з дескрипторами на інших зображеннях. В даному алгоритмі розпізнавання використовуються методи SIFT та SURF разом для всіх ключових точок, що знайдені обома методами. Таким чином, в результаті застосування алгоритму SIFT буде отримано m ключових точок, а в результаті застосування SURF – k , то загальна кількість ключових точок буде дорівнювати розмірності об'єднання множин, отриманих за допомогою SIFT та SURF.

2.1.3 Модифікація критерію подібності Керенексі

В [6] був представлений підхід для порівняння двох зображень на основі вейвлетів Габора та успішно адаптований та інтегрований з SIFT в [7].

Керенексі запропонував під час порівняння використовувати зважену суму двох компонентів порівняння: загального значення співпадіння двох обличчя та кількості успішно співставлених точок.

Локальність порівняння, тобто порівняння точок спільних частин обличчя, забезпечується за допомогою введення порогового значення, яке фільтрує точки з різних регіонів.

Нехай T – тестове зображення, G – зображення з бази даних. Для кожної ключової точки $t \in T$ визначається множина релевантних ключових точок $g \in G$ за формулою 2.1:

$$\sqrt{(x_t - x_g)^2 + (y_t - y_g)^2} \leq \theta, \quad (2.1)$$

де x та y – координати ключових точок.

Якщо жодної релевантної точки не знайдено, точка t виключається з процедури порівняння.

Як зазначено в [8], неможливо обчислити оптимальне порогове значення аналітично, адже воно залежить від розмірів та якості зображення. Наприклад, для бази даних СТК це значення варіюється в межах [10,40].

Загальне співпадіння двох обличч OS обчислюється як середнє значення співпадінь усіх відповідних точок:

$$OS_{T,G} = mean\{S(t, g), t \in T, g \in G\}, \quad (2.2)$$

де S – критерій подібності косинусів.

Після цього визначається зображення з бази даних з найбільш подібною ключовою точкою до кожної ключової точки тестового зображення. Далі виконується обчислення наступного виразу:

$$G_i = \frac{C_i}{N_i}, \quad (2.3)$$

де C_i – кількість ключових точок i -го зображення з бази даних, які є найближчими до ключових точок тестового зображення;

N_i – загальна кількість ключових точок в i -му зображенні з бази даних.

Зважена сума двох критеріїв подібності обчислюється за наступною формулою:

$$FS_{t,i} = \alpha OS_{t,i} + \beta G_i, \quad (2.4)$$

де t – тестове зображення;

i – індекс зображення з бази даних.

Індекс зображення з БД, зважена сума якого є максимальною серед всіх зображень в БД, є результатом розпізнавання тестового обличчя. [7]

Не зважаючи на високі показники розпізнавання [7], є два пункти, які роблять даний підхід не зовсім гнучким для вирішення певних задач, а саме:

- використання порогового значення, яке визначається лише емпірично;
- не нормалізоване значення критерію 2.4 та коефіцієнтів.

Для обчислення порогового значення необхідно мати тестовий набір даних, що є не зовсім можливо в деяких задачах класу автоматичного розпізнавання облич, а також значно ускладнюється процес підготовки зображення до аналізу. У свою чергу, поточна область значень критерію 2.4 значно ускладнює вирішення задачі розпізнавання з можливістю використання варіанту не знаходження співпадінь, адже порогове значення тепер також необхідно визначати емпірично і відкритим постає питання чи можна взагалі адаптувати метод до таких умов.

Одним із методів заміни порогового значення є використання методу кластеризації k-середніх. Кількість кластерів можна встановити в п'ять (два ока, ніс, ліва та права нижні частини обличчя)[9].

Альтернативою, яка використовується в даній модифікації, є виявлення координат ключових точок обличчя (очей, носу, центру губ). Цього можна досягти за допомогою використання методів детектування (наприклад, алгоритму Віоли-Джонса) на тестових базах даних, які є у вільному доступі. Після успішного виявлення координат (центрів кластерів), співвіднести кожну ключову точку з одним з кластерів, Евклідова відстань до якого буде найменшою.

Для нормалізації критерію 2.4 пропонується замість зваженої суми критеріїв обчислювати їх середнє арифметичне (або середнє зважене арифметичне, якщо необхідно вводити вагові коефіцієнти). У випадку відсутності необхідності вводити вагові коефіцієнти, вираз матиме наступний вигляд:

$$FS_{t,i} = mean(OS_{t,i}, G_i) \quad (2.5)$$

Така зміна виразу 2.4 вводить більш зрозумілу метрику обчислення подібності облич на зображеннях – процентну, тому можна встановити більш зрозуміле порогове значення і якщо не знайдено жодного значення, яке його перевищує, то вважати, що співпадінь в базі даних не знайдено.

2.1.4 Модифікація алгоритму Брона-Кербоша для ідентифікації осіб

Сам процес ідентифікації осіб серед набору зображень облич можна представити як поєднання схожих зображень за критерієм, описаним в п.2.1.3. Порівнявши всі зображення, отримаємо граф з великою степенню щільності. Цю степінь можна зменшити за допомогою введення порогового значення, нижче якого зв'язок між зображеннями ігнорується. В отриманому графі необхідно виділити підграфи таким чином, що кількість зв'язків між вершинами буде якомога більшою і водночас середнє арифметичне значення подібності вершин не буде нижчим за порогове.

Як було зазначено в першому розділі, алгоритм Брона-Кербоша та його модифікації призначені для знаходження максимальних клік в графі, тобто повних підграфів, які не можуть бути розширені за допомогою включення додаткової вершини без втрати властивості повного графу (кожна пара вершин має ребро). Таким чином, використавши цей алгоритм, обидві умови (максимальна кількість зв'язків та подібність – вища за порогову) будуть виконані. Недоліками використання даного методу є:

- велика асимптотична складність алгоритму;
- відсутність механізму для подолання аномалій. Наприклад, вершина, що має зв'язки з усіма вершинами підграфу, окрім одної, не буде додана до нього;
- знайдені кліки можуть мати спільні вершини, що є нелогічним з точки зору інтерпретації результатів роботи алгоритму – одне зображення обличчя може належати декільком особам.

Для вирішення цих проблем пропонується підхід, що складається з наступних кроків:

- розбити граф на підграфи меншого розміру;
- відмітити ребра підграфу, що відповідають заданим критеріям (наприклад, степінь схожості вершин більша за порогове значення);
- запустити модифікацію алгоритму Брона-Кербоша з опорною вершиною на кожному підграфові. Результатом є список клік, що можуть мати спільні вершини;
- видалити спільні вершини із клік;
- побудувати повний граф клік, вершини якого матимуть спільне ребро з наступними атрибутами: сумарна подібність та кількість зв'язків між елементами клік, подібність яких більша за порогове значення. Іншими словами, атрибути даного ребра агрегують значення зв'язків ребер між вершинами даної кліки та іншої;
- відмітити ребра графу, що відповідають заданим критеріям (наприклад, значення подібності всіх ребер, що з'єднують вершини різних клік, мають бути більшими за порогове значення);
- запустити модифікацію алгоритму Брона-Кербоша з опорною вершиною на графові клік;
- видалити спільні вершини з результуючих клік.

Процес видалення вершин з клік складається з наступних кроків:

- вибір вершини зі списку дублікатів;
- визначення кліки, в якій необхідно залишити вершину;
- видалити вершину з інших клік.

Ці дії необхідно виконати для всіх вершин, що дублюються.

Згідно із загальним методом та процесом видалення сумісних вершин, описаних вище, в модифікованому алгоритмі є декілька місць, налаштувавши які, можна отримати специфічні результати роботи алгоритму в цілому:

- використавши різні стратегії відбору зв'язків між вершинами графу, можна отримати кліки, що складаються з різних вершин;
- використати різні стратегії вибору вершини зі списку дублікатів;

- використати різні стратегії вибору клік, в який необхідно залишити вершини, що дублюються.

Так як, фактично, ми послідовно використовуємо модифікацію алгоритму спочатку на підграфах графу, а потім на результуючому графові клік, точки модифікації є актуальними як для початкових підграфів, так і для графів клік.

В дисертації для початкових підграфів використовуються наступні стратегії:

- критерій відбору зв'язків – степінь подібності є більшою за порогове значення;
- стратегія вибору вершини зі списку дублікатів – вершина має найбільшу кількість зв'язків, відібраних за стратегією, описаною вище;
- стратегія вибору кліки – найвища степінь подібності між вершиною-дублікатом та вершиною, що не має дублікатів.

Для графу клік використовуються наступні стратегії:

- критерій відбору зв'язків – співвідношення кількості зв'язків, степінь подібності яких вища за порогову, до загальної кількості зв'язків – більша за порогове значення (від 0 до 1);
- стратегія вибору вершини зі списку дублікатів – сума степенів подібності є найвищою. Іншими словами, обрана вершина найбільше впливає на середню степінь подібності між кліками;
- стратегія вибору кліки – найвища сумарна подібність.

В результаті роботи модифікованого алгоритму будуть отримані підграфи графу порівняння зображень облич, в яких є повні підграфи, а співвідношення кількості зв'язків між вершинами різних підграфів, степінь подібності яких вища за порогову, до загальної кількості зв'язків – більша за порогове значення.

Використання даної модифікації повністю або частково вирішує недоліки алгоритму Брона-Кербоша та його модифікацій:

- в середньому, асимптотична складність буде зменшена, адже складність алгоритму Брона-Кербоша з опорною вершиною має складність $O(3^{\frac{n}{3}})$, а запропонована модифікація - $kO(3^{\frac{n}{3k}}) + O(3^{\frac{m}{3}})$, де k – кількість, на яку розбивається основний граф перед початком роботи запропонованого алгоритму, а m – кількість клік, отриманих в результаті виконання алгоритму Брона-Кербоша з опорною вершиною на підграфах. В найгіршому випадку асимптотична складність зменшена не буде;
- з'явилась можливість подолати аномалії за рахунок введення критерію відбору зв'язків клік;
- результатом роботи є підграфи, що не мають сумісних вершин.

2.2 Вирішення задачі напівавтоматичної класифікації осіб

2.2.1 Постановка задачі

Існує список користувачів соціальних мереж, що класифіковані наступним чином:

- приймав участь, сторона незаконних бандформувань;
- приймав участь, український військовий;
- не приймав участь у конфлікті.

Виконати класифікацію користувачів, що присутні на одних фотографіях з класифікованими. У випадку присутності на одних фотографіях з користувачами різних класів, пріоритет класифікації вважати наступним: сторона незаконних бандформувань, український військовий, не приймав участь у конфлікті.

2.2.2 Опис запропонованого рішення

Так як в рамках магістерської дисертації основними вхідними даними для класифікації є зображення з профілів користувачів соціальних мереж, головним критерієм для класифікації є присутність особи, що класифікується, з особою, профіль користувача якої в соціальній мережі був класифікований в ручному режимі.

Таким чином, процес класифікації складається з наступних кроків:

- за результатами автоматичної ідентифікації (п.2.1) необхідно визначити, на яких фотографіях які саме особи ідентифіковані разом;
- встановлення зв'язку між профілем користувача в соціальній мережі та особою, ідентифікованою в результаті автоматичної ідентифікації;
- класифікація користувачів, які присутні на одній фотографії з користувачами, які були класифіковані в ручному режимі.

В рамках дисертації використовується наступний критерій відповідності між профілем користувача та ідентифікованою особою: особа присутня на більшості фотографій, що належать профілю користувача. У випадку, якщо одна особа присутня на більшості фотографій в декількох профілях, обирається профіль з найбільшою кількістю фотографій особи.

Висновки до розділу

В розділі запропоновано методи для вирішення задач автоматичної ідентифікації осіб на зображеннях та їх напівавтоматичної класифікації.

В рамках вирішення задачі автоматичної ідентифікації запропонована модифікація критерію подібності зображень Керепескі, результатом роботи якої є нормалізоване значення подібності двох зображень.

Використання запропонованої модифікації алгоритму Брона-Кербоша в середньому зменшує його асимптотичну складність, дозволяє ігнорувати аномалії у вхідних даних, а результатом роботи є підграфи, що не мають спільних вершин, з заданими характеристиками.

Для вирішення задачі напівавтоматичної класифікації був запропонований метод, що базується на результатах роботи алгоритму групування схожих зображень.

РОЗДІЛ 3 РЕАЛІЗАЦІЯ СИСТЕМИ

3.1 Постановка задачі

Розробити систему, яка виконуватиме класифікацію осіб за зображеннями, отриманими із соціальних мереж. Вона має відповідати наступним критеріям:

- система повинна мати мікросервісну архітектуру;
- в компонентах системи мають бути реалізовані методи, описані в розділі 2;
- обробка даних має бути виконана за допомогою технологій паралельних та розподілених обчислень;
- компоненти системи мають бути слабко зв'язані між собою;
- проміжні дані та результати класифікації мають бути збережені в персистентному сховищі даних.

3.2 Контекст роботи системи

Система, що була розроблена, є частиною платформи для класифікації осіб за даними соціальних мереж, яка, у свою чергу, складається з наступних систем:

Система класифікації осіб, що виконує підготовку отриманих даних із системи збору даних і, власне, класифікує осіб;

система збору даних про користувачів соціальних мереж, що збирає дані про користувачів, які є наявними у відкритому доступі в соціальних мережах, з якими вона взаємодіє за допомогою SDK або API. Також вона виконує періодичне оновлення завантажених даних.

В процесі роботи платформи можна виділити декілька етапів:

- оператор системи класифікації осіб встановлює початкові дані та ініціює процес класифікації;
- Система класифікації осіб ініціює процес збору даних про користувача системи збору даних;

- Система збору даних, використовуючи SDK або публічне API соціальної мережі, отримує необроблені дані користувача та зберігає їх у власному сховищі;
- Система збору даних повідомляє системі класифікації про завершення підготовки даних;
- Система класифікації отримує необроблені дані, виконує їх обробку та класифікацію користувача, після чого ініціює процес класифікації його друзів;
- оператор системи класифікації в будь-який момент може отримати поточні результати класифікації.

Схематично складові частини платформи, а також їх взаємодія зображена на рисунку 3.1.

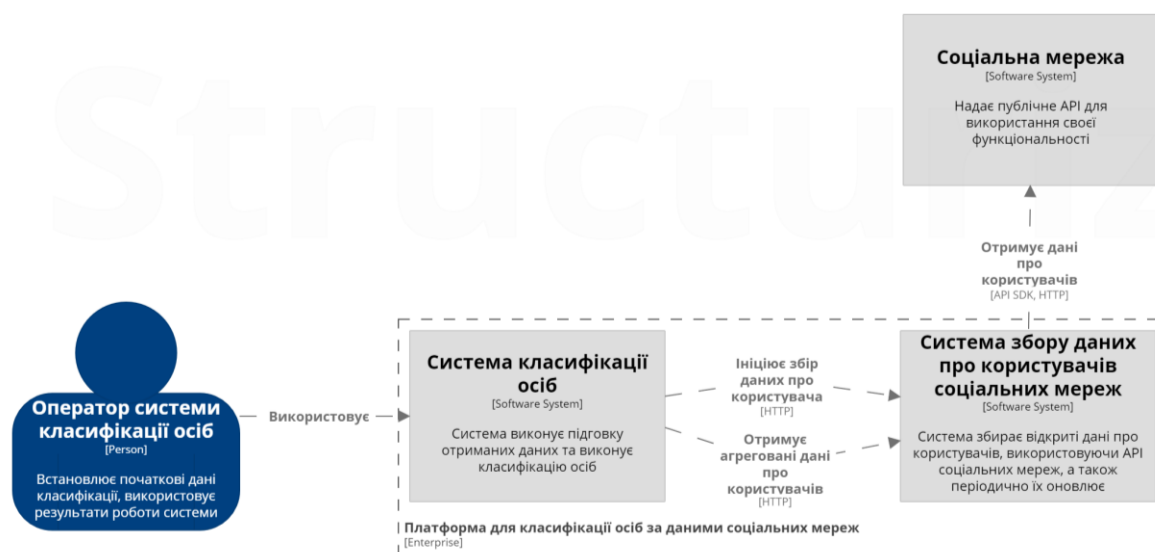


Рисунок 3.1 – Діаграма контексту платформи для класифікації осіб за даними соціальних мереж

В рамках магістерської дисертації була реалізована система класифікації осіб. В рамках системи збору даних про користувачів соціальних мереж було спроектовано пу-

блічне API та реалізована макетна система, яка конфігурується даними про користувачів, необхідних для роботи системи класифікації, та імітує результат роботи з соціальними мережами.

3.3 Опис складових частин системи

Система класифікації осіб має мікросервісну архітектуру і складається з наступних сервісів:

- Супервізор - вхідна точка системи, координує роботу всіх компонентів, відслідковує прогрес класифікації кожного користувача;
- Менеджер даних - надає API для взаємодії з підготовленими даними та результатами класифікації;
- Конвеєр підготовки даних - виконує підготовку агрегованих даних про користувача із системи збору даних для подальшого аналізу та надсилає їх менеджеру даних для збереження;
- Конвеєр обчислення степені схожості користувачів - виконує обчислення степені схожості користувачів за підготовленими даними та наперед зазначеними критеріями;
- Конвеєр групування користувачів - на основі степені схожості користувачів виконує їх групування;
- Конвеєр класифікації користувачів - на основі даних групування, а також початкових даних класифікації деяких користувачів, виконує класифікацію осіб;
- Брокер повідомлень – зберігає повідомлення в темах, організовує підписки та дозволяє зчитати повідомлення з теми. Слугує як компонент взаємодії сервісів через інтерфейс передачі сигналів та повідомлень, що дозволяє розірвати безпосередній зв'язок між сервісами.

Етапи процесу класифікації користувача із залученням сервісів, описаних вище, схематично зображено в Додатку А. З метою спрощення діаграми процес передачі управління між сервісами (лінії взаємодії відмічені синім кольором) схематично зображені в додатку Б.

Процес класифікації користувача складається з наступних етапів:

- користувач системи ініціює запуск процесу обробки, передавши його ідентифікатор в соціальній мережі Супервізору;
- Супервізор звертається до Менеджеру даних з метою отримати поточний стан класифікації користувача. Можливі наступні варіанти:

- 1) дані користувача – відсутні;
- 2) дані користувача – підготовлені і готові до обчислення степені схожості;
- 3) дані користувача – підготовлені і степінь схожості обчислена з іншими користувачами;

- на основі варіантів готовності, згаданих вище, Супервізор виконує одну з наступних дій:

- 1) надсилає повідомлення Брокерові повідомлень до теми `data_preparation`, якщо дані користувача – відсутні;
- 2) надсилає повідомлення Брокерові повідомлень до теми `data_comparison`, якщо дані користувача – підготовлені і готові до обчислення степені схожості;
- 3) ігнорує запит, якщо степінь схожості даних користувача була обчислена;

- отримавши повідомлення з теми `data_preparation` Брокера повідомлень, Конвеєр підготовки даних розпочинає процес підготовки, який складається з декількох етапів:

- 1) отримання даних користувача з Системи збору даних про користувачів соціальних мереж;

- 2) виконання підготовки даних;
- 3) збереження підготовлених даних в Менеджері даних;
- 4) ініціалізація процесу порівняння даних користувача з іншими користувачами через надсилання повідомлення Брокерові повідомлень до теми `data_comparison`;

- отримавши повідомлення з теми `data_comparison` Брокера повідомлень, Конверсер обчислення степені схожості користувачів розпочинає процес порівняння користувачів, який складається з декількох етапів:

- 1) отримання підготовлених даних до порівняння з Менеджеру даних;
- 2) виконання порівняння даних поточного користувача з іншими користувачами;
- 3) збереження даних порівняння в Менеджері даних;
- 4) ініціалізація процесу групування даних користувача з іншими користувачами через надсилання повідомлення Брокерові повідомлень до теми `data_grouping`;

- отримавши повідомлення з теми `data_grouping` Брокеру повідомлень, Конверсер групування користувачів розпочинає процес групування даних поточного користувача, яких складається з декількох етапів:

- 1) отримання результатів порівняння даних користувача з іншими користувачами з Менеджеру даних;
- 2) групування даних поточного користувача, а також виконання обчислень, необхідних для процесу загального групування;
- 3) збереження результатів групування в Менеджері даних;
- 4) ініціалізація процесу класифікації користувача через надсилання повідомлення Брокерові повідомлень до теми `user_classification`;

- отримавши повідомлення з теми `user_classification` Брокеру повідомлень, Конвеєр класифікації користувачів розпочинає процес класифікації користувача, який складається з декількох етапів:

- 1) отримання результатів групування даних користувача з іншими користувачами з Менеджеру даних;
- 2) виконання проміжної класифікації користувача;
- 3) збереження результатів проміжної класифікації в Менеджері даних;
- 4) ініціалізація процесу класифікації друзів користувача за допомогою Супервізора;

Реалізація системи на мікросервісній архітектурі дає наступні переваги окремим категоріям людей:

- ІТ департаменту:

- 1) розділення кодової бази спрощує процес тестування та внесення змін, адже замість одного сервісу на десятки тисяч компонентів є сотня сервісів по сто компонентів;
- 2) компоненти можуть бути поділені між командами розробників, що чітко розмежує границі відповідальності за працездатність системи в цілому;
- 3) можливість більш гнучко налаштовувати інфраструктуру, що дозволить зменшити об'єм витрат на неї, підвищити відмовостійкість системи в цілому;

- бізнесу:

- 1) збільшення швидкості внесення змін до бізнес логіки системи, адже замість проходження процесу CI/CD для всієї системи цей процес буде пройдено лише для частини сервісів, в яких було виконано зміни;
- 2) зменшення часу релізних робіт, в рамках яких система є повністю бездієвою.

3.4 Опис Супервізора

Супервізор є фасадом системи класифікації осіб. Він надає публічне API користувачеві, яке складається з наступних функцій:

- ініціалізація процесу класифікації користувача;
- класифікація користувача в ручному режимі. Використовується для задання початкових даних класифікації, на основі яких будуть класифіковані інші користувачі;
- отримання результатів класифікації.

Супервізор виконує функцію моніторингу процесу класифікації по кожному користувачеві в системі, тому надає API іншим сервісам для реєстрації початку та кінця етапу класифікації користувача.

Іншою функцією Супервізора є координація роботи всіх сервісів системи. Отримавши запит на ініціалізацію процесу класифікації користувача, він вирішує, з якого саме етапу необхідно почати цей процес – з підготовки даних чи їх порівняння, і чи взагалі є в цьому необхідність, адже користувач може вже бути класифікованим на цей момент. Такий підхід дозволяє зменшити витрати обчислювальних ресурсів та спростити механізми зберігання цілісності даних, адже користувач в системі, а також його дані, залишаться унікальними і не буде дублювання результатів порівняння, групування та класифікації.

Не менш важливим є те, що Супервізор може перервати подальший процес обробки користувача. Розглянемо наступну ситуацію:

- був отриманий запит на класифікацію користувача А;
- процес класифікації був успішно завершений;
- система автоматично ініціює процеси класифікації усіх його друзів – В, С;
- процес класифікації користувача В був успішно завершений;
- система автоматично ініціює процеси класифікації усіх його друзів – А, С;
- Супервізор відхиляє запит на ініціацію процесів, тому що користувач А був класифікований, а для С процес був ініційований.

Окрім оптимізації витрат ресурсів та збереження цілісності, ця функція Супервізора дозволяє зупинити процес класифікації, коли всі користувачі будуть класифіковані.

Супервізор безпосередньо взаємодіє з Брокером повідомлень та Менеджером даних, а також використовується іншими сервісами для моніторингу.

Сервіс реалізовано з використанням технології Spring Boot.

3.5 Опис Конвеєра підготовки даних

Конвеєр підготовки даних виконує попередню обробку агрегованих даних із Системи збору даних про користувачів соціальних мереж, трансформує їх у необхідний для порівняння формат та надсилає Менеджерові даних для збереження.

Так як основний формат даних, який використовує система, є зображення, попередня обробка має наступні етапи:

- розпізнавання облич за допомогою інтеграції з сервісом Google Cloud Vision API;
- знайдені обличчя вирізаються з зображень, формуючи нові зображення;
- зображення переводяться в напівтоновий формат колориметричним методом;
- виконується процес еквалізації гістограми зображень;
- нормалізуються координати положення центрових точок (центроїдів), таких як ліве та праве око, ніс, центр губ, підборіддя.

Трансформація попередньо оброблених даних у необхідний для порівняння формат включає в себе наступні етапи:

- 1) знаходження ключових точок на зображеннях облич за допомогою методів SIFT та SURF;
- 2) знайдені ключові точки співвідносяться до одного з центроїдів.

З результату трансформації і оброблених даних, а також метаданих (ідентифікатор фотографії та альбому, на яких були знайдені обличчя, а також ідентифікатор власника фотографій) формується документ та надсилається Менеджеру даних для збереження.

З метою прискорення роботи конвеєра, адже кількість користувачів та їх фотографій може сягати мільйонів одиниць, було прийняте рішення використати технології обробки надвеликих масивів даних, а саме Apache Beam, що є аналогом Apache Spark та відрізняється тим, що надає уніфікований інтерфейс для технологій Big Data, однією з реалізацій якої є Apache Spark. Так як система використовує окремі технології Google Cloud, реалізацією інтерфейсу є Google Cloud Data Flow. Основною перевагою використання технологій Big Data є горизонтальне масштабування обчислювальних потужностей з метою збільшення швидкості обробки даних, а використання для цього Apache Beam дозволяє абстрагуватися від конкретної реалізації технології розподілених обчислень.

Для знаходження ключових точок на зображеннях обличч використовується бібліотека JavaCV, яка є адаптованою бібліотекою OpenCV на JVM.

Конвеєр запускається в автоматичному режимові за розкладом, вираз якого знаходиться в конфігураційних файлах сервісу. У випадку, якщо на момент запуску конвеєру, попередній не закінчив свою роботу, новий конвеєр не запускається.

3.6 Опис Конвеєра обчислення степені схожості користувачів

На основі даних, підготовлених Конвеєром підготовки даних, сервіс виконує порівняння зображень обличч згідно з підходом, описаним в п.2.1.2.

Отримавши повідомлення про необхідність порівняння зображень обличч, знайдених на фотографія профілів двох користувачів, конвеєр формує завдання з наступним принципом: всі зображення обличч, знайдені на фотографіях першого користувача мають бути порівняні з усіма зображеннями обличч, знайденими на фотографіях другого користувача. Окремим випадком є отримання повідомлення, в якому ідентифікатор першого

користувача дорівнює ідентифікатору другого. Іншими словами, необхідно виконати порівняння облич на фотографіях, що належать одному користувачеві. Дійсно, на фотографіях користувача можуть бути знайдені різні особи, тому необхідно однозначно ідентифікувати їх усіх. Для цього необхідно виконати порівняння зображень облич. Однак в цьому випадку, важливо розуміти, чи два поточні обличчя знайдені на одній фотографії, чи на різних. Важко уявити ситуацію, що на одній фотографії присутня одна і та ж особа декілька разів (за виключенням ситуації, коли є відображення особи в дзеркалі, в рамках дисертації цей випадок ігнорується). У зв'язку з цим, виконується порівняння зображень облич, що присутні на різних фотографіях.

Результатом роботи конвеєру є список результатів завдань, кожне з яких складається із самого завдання, тобто двох ідентифікаторів облич, і результату порівняння, тобто процент схожості облич. Цей список надсилається Менеджерові даних для збереження.

Конвеєр також реалізовано за допомогою технологій Big Data, а саме з використанням Apache Spark.

Конвеєр запускається в автоматичному режимові за розкладом, вираз якого знаходиться в конфігураційних файлах сервісу. У випадку, якщо на момент запуску конвеєру, попередній не закінчив свою роботу, новий конвеєр не запускається.

3.7 Опис Конвеєра з групування користувачів

Конвеєр з групування користувачів інкапсулює логіку автоматичного методу ідентифікації осіб на зображеннях, за допомогою якого була вирішена задача, поставлена в п.2.1.4 роботи. Складається з двох конвеєрів: потокового та пакетного, а також допоміжних компонентів:

- Компонент інтеграції з менеджером даних, який використовує API менеджера даних для отримання результатів порівняння облич, розпізнаних на зображеннях з профілів користувачів, а також збереження результатів групування;

- Компонент інтеграції з супервізором – повідомляє Супервізор про початок та кінець етапу групування користувачів;
- Тригер конвеєру – компонент, що активує потоковий та пакетний конвеєри групування користувачів за розкладом;
- Компонент інтеграції з брокером повідомлень – надсилає та отримує повідомлення з брокеру повідомлень, виконує валідації та конвертацію повідомлень у формат, необхідний конвеєрам для правильної роботи;
- Потоковий конвеєр – виконує групування облич на зображеннях за результатами порівняння облич, Виконаними конвеєром обчислення степені схожості користувачів;
- Пакетний конвеєр – виконує групування груп облич, знайдених поточним конвеєром.

Процес поточного групування складається з наступних кроків:

- Тригер конвеєру розпочинає процес поточного групування;
- Конвеєр поточного групування сповіщає Супервізор через Компонент інтеграції з Супервізором про початок поточного групування облич, детектованих на зображеннях поточного користувача (якщо в повідомленні було зазначено, що необхідно виконати дану операцію в рамках одного користувача) або підготовки до пакетного групування двох користувачів;
- Конвеєр поточного групування зчитує повідомлення, додані до теми `data_grouping` за допомогою компоненту інтеграції з Брокером повідомлень;
- Компонент інтеграції з Брокером повідомлень виконує валідацію отриманих повідомлень, а також конвертацію валідних повідомлень до формату, придатному до обробки Поточним конвеєром;
- Конвеєр поточного групування надсилає запит Компонентів інтеграції з Менеджером даних на отримання даних порівняння облич на зображеннях поточних користувачів;

- Компонент інтеграції з Менеджером даних отримує дані порівняння облич на зображеннях поточних користувачів та передає їх Конвеєру потокового групування;
- Компонент потокового групування виконує групування облич користувачів за алгоритмом, наведеним в п. 2.1.4, та передає результати групування Компоненту інтеграції з Менеджером даних, який їх зберігає в Менеджері даних;
- Конвеєр потокового групування надсилає запит Компоненту інтеграції з Супервізором про завершення етапу групування зображень користувача або підготовки поточних користувачів до пакетного групування, який передає цю інформацію Супервізору;
- Конвеєр потокового групування передає дані Компоненту інтеграції з Брокером повідомлень для надсилання повідомлень про необхідність виконання класифікації користувачів, який, у свою чергу, надсилає повідомлення Брокерові повідомлень до теми `user_classification`.

Процес пакетного групування складається з наступних етапів:

- Тригер конвеєру розпочинає процес пакетного групування;
- Пакетний конвеєр надсилає запит Компоненту інтеграції з менеджером даних на отримання результатів групування Потоковим конвеєром усіх користувачів, який, у свою чергу, отримує ці дані з Менеджеру даних;
- Пакетний конвеєр виконує процес групування груп, сформованих Потоковим конвеєром, отримуючи ідентифікованих осіб на зображеннях, отриманих із соціальних мереж;
- Пакетний конвеєр передає ідентифікованих осіб (згруповані групи) Компонентові інтеграції з менеджером даних, який, у свою чергу, зберігає ідентифікованих осіб в Менеджері даних.

Структура Конвеєра групування користувачів, а також зв'язки між компонентами зображені схематично в додатку В.

3.8 Опис Конвеєра класифікації користувачів

Конвеєр класифікації користувачів інкапсулює метод класифікації, запропонований в п.2.2 роботи. Він використовує результати групування, отримані в результаті роботи Конвеєру групування користувачів.

Даний компонент працює лише в режимі пакетної класифікації, тобто виконується класифікація всіх користувачів одночасно. В потоковому режимі компонент ініціює процес класифікації друзів поточного користувача, надіславши відповідні запити Супервізору. Результатом роботи є маркування вузлів DetectedUser та User в Neo4J одним із наступних значень: COLLABORATOR, WARRIOR, UNCLASSIFIED.

3.9. Опис Менеджера даних

Менеджер даних виконує наступні функції в системі:

- інкапсуляція низькорівневої логіки взаємодії з персистентними сховищами даних за високорівневим інтерфейсом;
- абстрагування сервісів системи від конкретних реалізацій сховищ даних, що дає можливість змінювати тип сховища даних з метою збільшення продуктивності системи в цілому;
- спрощення процесу забезпечення цілісності даних в сховищах даних, адже лише один сервіс знає про реальну структуру даних, їх організацію та механізм транзакційного доступу.

Враховуючи структуру даних, які необхідно зберігати та опрацьовувати, було прийняте рішення зберігати взаємозв'язки між користувачами, альбомами, фотографіями та обличчями, знайдених на фотографіях, а також результат їх групування в графовій базі даних Neo4j, яка реалізована для потреб в моделюванні даних, які можуть бути представлені у вигляді великих графів і надає зручну мову виконання запитів Cypher. Наприклад, в порівнянні із звичайною реляційною базою даних, деякі типи запитів мо-

жуть бути виконані на Neo4J за константний час в той час, як результати роботи реляційної бази даних не будуть отримані взагалі за розумний час через експоненційний вибух. Також такі запити будуть з легкістю сформульовані за допомогою синтаксису Cypher в порівнянні з десятками чи навіть сотнями джоїнів на мові SQL.

Результати підготовки даних Конвеєром підготовки даних, у свою чергу, найдоцільніше зберігати в документ-орієнтованих базах даних через наступні причини:

- структура даних – документ з вкладеними композитними елементами;
- необхідність отримати цілий документ, адже взаємодія з частиною документа не має сенсу.

В якості реалізації даного типу бази даних була обрана MongoDB.

Структура документу в MongoDB наведена в таблиці 3.1.

Таблиця 3.1 – Опис атрибутів типу FaceDescriptor

| Назва атрибуту | Тип | Опис атрибуту |
|--------------------|--------------|--|
| faceId | String | Ідентифікатор обличчя (первинний ключ) |
| photoOwnerId | UUID | Ідентифікатор користувача в соціальній мережі |
| albumId | UUID | Ідентифікатор альбому користувача |
| photoId | UUID | Ідентифікатор фотографії користувача, на якій було детектовано обличчя |
| photoUrl | String | URL фотографії в соціальній мережі |
| rawImage | String | Зображення обличчя, заповнене у форматі Base64 |
| groupedDescriptors | FaceCentroid | SIFT+SURF дескриптори ключових точок навколо центроїду (очей, носа, центру губ і т.д.) обличчя |
| boundaries | FaceKeypoint | Границі зображення обличчя на фотографії |

Структура частини документу FaceCentroid наведена в таблиці 3.2.

Таблиця 3.2 – Опис атрибутів типу FaceCentroid

| Назва поля | Тип | Примітка |
|---------------------------|--------------------|--|
| landmark | Landmark | Тип центроїду (ліве/праве око, центр носу, підборіддя та губ) |
| center | FaceKeypoint | Координати центроїду |
| cen- troidD'escriptors | Map<String,String> | Мапа, ключем якої є координати ключової точки у форматі "x y", а значенням – SIFT+SURF дескриптори ключової точки, закодовані у форматі Base64 |

Структура частин документу FaceKeypoint наведена в таблиці 3.3.

Таблиця 3.3 – Опис атрибутів типу FaceKeypoint

| Назва поля | Тип | Примітка |
|------------|--------|-----------------------------|
| x | String | Координата x ключової точки |
| y | String | Координата y ключової точки |

Структурна схема графової бази даних зображена на рисунку 3.2.

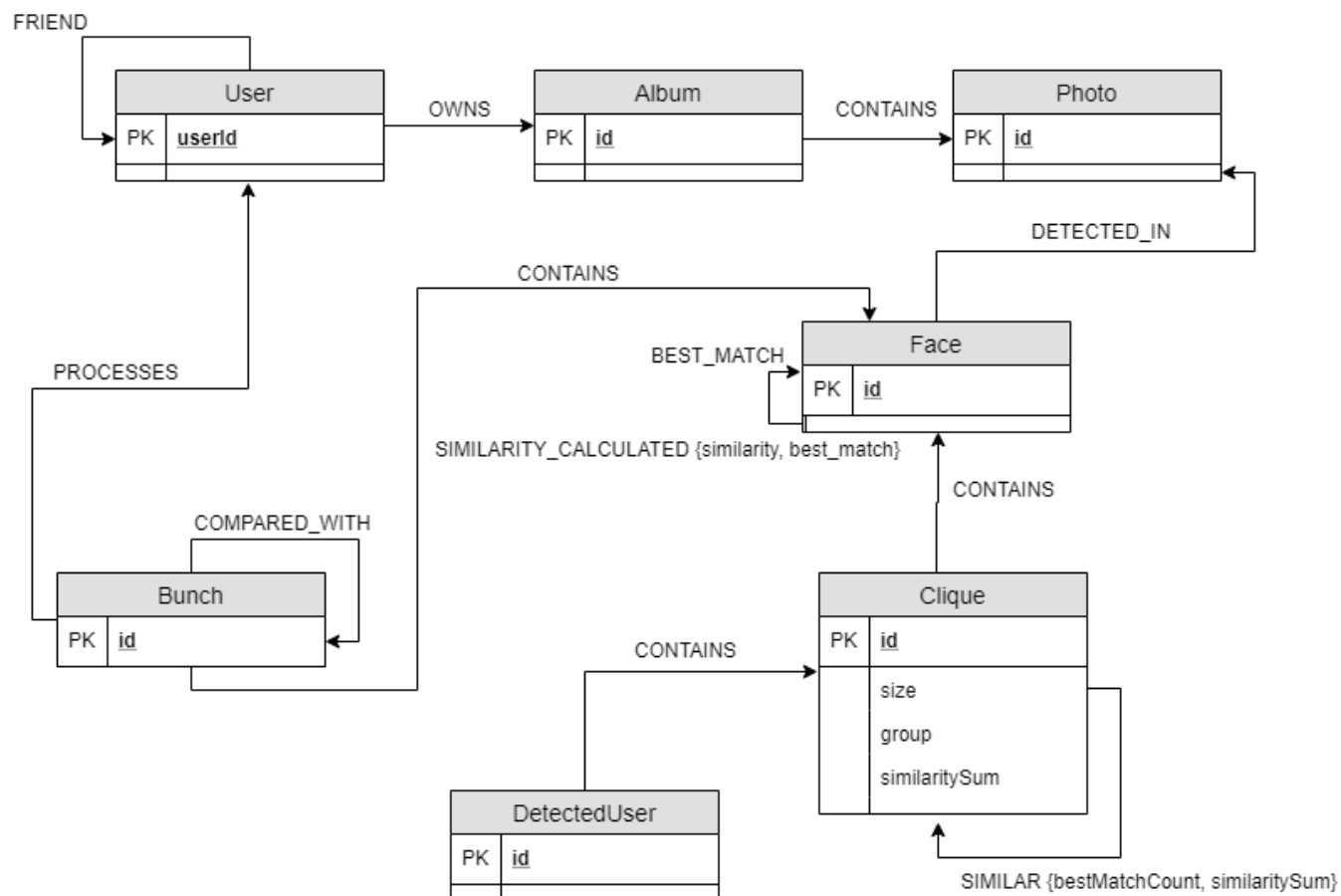


Рисунок 3.2 – Структурна схема графової бази даних

Структурна схема графової бази даних складається вершин наступних типів:

- User – користувач соціальних мереж;
- Album – альбом користувача соціальних мереж;
- Photo – фотографія, що знаходиться в альбомі користувача;
- Face – обличчя, розпізнане на фотографії;
- Bunch – пакет даних користувача;
- Clique – відображає групу облич в межах одного користувача, які задовольняють критеріям схожості. Іншими словами, розпізнана особа на фотографіях користувача;
- DetectedUser – відображає групу Clique в базі даних. Іншими словами, розпізнана особа серед усіх користувачів.

Висновки до розділу

Розроблено систему, яка виконує класифікацію осіб за зображеннями, отриманими із соціальних мереж. Вона відповідає наступним критеріям:

- мікросервісна архітектура;
- в компонентах системи реалізовані методи, описані в розділі 2;
- обробка даних виконана за допомогою технологій Big Data;
- забезпечена слабка зв'язність компонентів завдяки використанню Брокеру

повідомлень;

- так як дані в системі зручно зберігати у двох різних форматах – документу та графовій структурі, були використані бази даних, які призначені для роботи саме з такими форматами даних.

РОЗДІЛ 4 ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ

4.1 Загальні відомості

В якості датасету була обрана база даних ORL, яка складається з зображень 40 людей (по 10 зображень кожної). Зображення містять лише обличчя та мають однотонний фон, в яких змінюється розміри, кути нахилу голови, умови освітлення. Розміри: 92x112 пікселів.

Дослідження ефективності розпізнавання облич проводились на базі даних, підготовлених системою. Так як класифікатор встановлюватиме відповідність між особами, розпізнаними на зображеннях, та користувачами соціальної мережі за критерієм максимальності розміру групи, розпізнаної в рамках користувача, було прийняте рішення додавати не всі зображення користувачам випадковим чином, а використовуючи наступну пропорцію: 70% зображень залишати в рамках кожного користувача, 30% – додавати іншим користувачам випадковим чином до існуючих зображень та інтерпретувати їх як пари зображень. Даний підхід дозволяє збільшити стабільність процесу встановлення відповідності між особою та користувачем і є лінійним, а тому його точність буде 100% завжди, тому що це чисто технічна операція. Також цей спосіб розподілу дозволяє підготувати експериментальні дані для перевірки групування зображень користувачів на другому етапі (групування, виконаного на графові клік) і встановити семантичні зв'язки присутності на одному зображенні між користувачами.

4.2 Дослідження ефективності модифікованого критерію подібності при порівнянні облич з базою даних облич

Одне зображення кожної людини використовується як базове, а інші 9 – тестові.

В якості детектору частин обличчя (очей, носу, губ) використовується Google Cloud Vision API.

Результати порівняння з методами, наведеними в [5] наведені в таблиці 4.1.

Таблиця 4.1 – Порівняння результатів класифікації на датасеті ORL

| Метод порівняння | Процент співпадіння |
|----------------------|---------------------|
| Aly | 61.25 |
| Lenc-Kral | 78.75 |
| Керепексі | 80.56 |
| Запропонований метод | 81.94 |

4.3 Дослідження ефективності методу автоматичної ідентифікації осіб

Ефективність роботи досліджувалась на базі даних, підготовлених системою, опис якої наведено в розділі 3.

Перший етап групування, а саме групування в межах підграфів, було виконано під час процесу підготовки даних. Кількість підграфів дорівнює кількості користувачів у системі (див. п.4.1).

Результати ефективності роботи алгоритму групування наведені в таблиці 4.2.

Таблиця 4.2 – Результати ефективності роботи алгоритму групування

| | | |
|--|-----|------|
| Загальна кількість зображень | 400 | 100% |
| Кількість зображень в групах | 280 | 70% |
| Кількість зображень, розподілених між користувачами | 120 | 30% |
| Кількість правильно згрупованих зображень на першому етапі | 196 | 70% |
| Кількість правильно згрупованих зображень на другому етапі | 72 | 60% |
| Загальна точність роботи | 268 | 67% |

4.4 Дослідження ефективності роботи методу класифікації

Так як метод класифікації використовує дані ідентифікації користувачів і визначає спочатку присутність осіб на одному зображенні з точністю, яка дорівнює точності групування зображень на другому етапі (60%), а також встановленню відповідності між ідентифікованою особою та користувачем соціальних мереж, точність якого в найгіршому

випадку дорівнюватиме точності групування зображень на першому етапі (70%). Таким чином, в найгіршому випадку точність роботи алгоритму на базі даних ORL дорівнюватиме 42%.

4.4 Відомості про швидкодію роботи системи

Швидкість обробки даних системою була обчислена за сумарним часом класифікації користувачів з 600 зображеннями облич. На обчислювальній машині з процесором Intel Core i7-8550U (1.8 ГГц, 4 фізичні ядра, 8 віртуальних потоків) та 16 ГБ оперативної пам'яті 1600 МГц швидкість обробки зображень дорівнює 1300 за хвилину (підготовка зображень та їх порівняння). Під час роботи системи всі компоненти, окрім Брокеру повідомлень, були розгорнуті локально.

Висновки до розділу

Запропонована модифікація критерію подібності двох зображень дає кращі результати порівняння, ніж критерій Керепексі.

Загальна точність роботи методу автоматичного розпізнавання осіб – 67%. Для покращення результатів необхідно збільшувати точність методів підготовки зображень до групування.

Загальна точність роботи методу класифікації в найгіршому випадку дорівнюватиме 42% і може бути збільшена за рахунок покращення точності розпізнавання осіб.

Швидкодія роботи системи є достатньою, враховуючи той факт, що її компоненти, які виконують найбільш складні обчислювальні операції, можуть бути горизонтально масштабовані.

РОЗДІЛ 5 РОЗРОЗБЛЕННЯ СТАРТАП-ПРОЕКТУ

5.1 Опис ідеї проекту

Опис ідеї стартап-проекту наведено в таблиці 5.1.

Таблиця 5.1 – Опис ідеї стартап-проекту

| Зміст ідеї | Напрямки застосування | Вигоди для користувача |
|--|---|--|
| Розробка системи, яка на основі даних соціальних мереж виконує класифікацію користувачів. Є базовою платформою для подальшого використання в задачах аналізу користувачів соціальних мереж | 1. Таргетована реклама | Додатковий механізм для визначення цільової аудиторії, платформа для застосування власних механізмів для визначення цільової аудиторії |
| | 2. Класифікація користувачів соціальних мереж | Є базовою платформою для вирішення задач класифікації користувачів |
| | 3. Правоохоронна діяльність | Визначення екстреміських настроїв, додатковий механізм для знаходження правопорушників та контррозвідувальної діяльності |

5.2 Технологічний аудит проекту

В таблиці 5.2 наведено результати аналізу технологічної здійсненності ідеї проекту

Таблиця 5.2 – Аналіз технологічної здійсненності ідеї проекту

| № п/п | Ідея проекту | Технології її реалізації | Наявність техноло- гій | Доступність техно- логій |
|----------|---|---|---|---|
| 1 | Розробка платформи як базового механізму для вирішення задач аналізу та класифікації на основі даних соціальних мереж | Стек Big Data | Наявні стеки технологій з відкритим програмним доступом (Apache License), а також платні аналоги (GCP, AWS) | Є безкоштовні та платні |
| | | Інфраструктура | Є можливість скористатися послугами IAAS та PAAS провайдерів, а також повністю створити інфраструктуру самостійно | доступні, платні, є безкоштовний період |
| 2 | Розробка механізму автоматичної ідентифікації осіб | Розпізнавання обличчя в умовах неконтрольованого середовища | Розпізнавання на основі аналізу локальних ознак. Існує необхідність в адаптації методів до умов та вхідних даних поточної | Набір технологій характеризується високим ступенем правильності розпізнавання за наявності тестових |

Продовження таблиці 5.2

| | | | | |
|---|--|--|--|--|
| | | | задачі. Необхідна попередня обробка вхідних даних для збільшення стабільності роботи алгоритму | даних, в умовах контрольованого середовища. В умовах неконтрольованого середовища та без тестових даних точність розпізнавання значно зменшується. Методи аналізу локальних ознак є запатентованими. |
| | | Класифікація за результатами ідентифікації | Готового алгоритму не існує, але легко реалізується | Готовий класифікатор – недоступний. Набір технологій для побудови класифікатора є у вільному доступі. Готова технологія не синтезована. |
| Обрана технологія реалізації проекту: Окремі технології зі стеку Big Data від Apache та Google Cloud, інфраструктура – GCP IAAS, Google Cloud Vision API для попередньої обробки зображень, власна модифікація методу розпізнавання зображень на базі методів SIFT/SURF, власна розробка методу класифікації. | | | | |

5.3 Аналіз ринкових можливостей запуску стартап-проекту

В таблиці 5.3 наведено попередню характеристику потенційного ринку стартап-проекту.

Таблиця 5.3 – Попередня характеристика потенційного ринку стартап-проекту

| № п/п | Показники стану ринку (найменування) | Характеристика |
|----------|--|---|
| 1 | Кількість головних гравців, од | 2-3 в сегменті реклами, відсутні в сегменті аналізу користувачів соц. мереж, 1-2 в правоохоронному сегменті України |
| 2 | Загальний обсяг продаж, грн/ум.од | Дані відсутні |
| 3 | Динаміка ринку (якісна оцінка) | Зростає |
| 4 | Наявність обмежень для входу (вказати характер обмежень) | Дані відсутні |
| 5 | Специфічні вимоги до стандартизації та сертифікації | Дані відсутні |
| 6 | Середня норма рентабельності в галузі (або по ринку), % | Дані відсутні |

В таблиці 5.4 визначені потенційні групи клієнтів, їх характеристики, а також наведено сформований перелік вимог до товару для кожної групи

Таблиця 5.4 – Характеристика потенційних клієнтів стартап-проекту

| № п/п | Потреба, що формує ринок | Цільова аудиторія (цільові сегменти ринку) | Вимоги споживачів до товару |
|----------|------------------------------------|--|---|
| 1 | Механізми для таргетованої реклами | Представники ринку електронної торгівлі | Вимоги до продукції: - зручне API; - стабільність роботи при піковому навантаженні; - висока швидкість обробки запиту; |

Продовження таблиці 5.4

| | | | |
|---|---|---|--|
| | | | <ul style="list-style-type: none"> - наявність документації. Вимоги до постачальника: <ul style="list-style-type: none"> - Дотримання GDPR; - супровід програмного забезпечення; - надання прикладу інтеграції з існуючими системами для таргетингу реклами. |
| 2 | Додатковий канал для збору даних правопорушників | Правоохоронні органи | Вимоги до продукції є аналогічними вимогам з пункту 1 даної таблиці, а також наявність методу класифікації правопорушників |
| 3 | Розробка платформи для проведення аналізу користувачів різного роду | Наукові організації, держстат та інші статистичні організації, соціологічні організації | Вимоги до продукції є аналогічними вимогам з пункту 1 |

Результати аналізу факторів, що перешкоджають ринковому впровадженню проекту, наведено в таблиці 5.5.

Таблиця 5.5 – Результати аналізу факторів загроз ринкового впровадження проекту

| № п/п | Фактор | Зміст загрози | Можлива реакція компанії |
|-------|------------|--|---|
| 1 | Політичний | Блокування окремих інтернет ресурсів на території України. | Використання різних PAAS / IAAS ресурсів для обходження блокування на технічному рівні. Використання VPN. |
| 2 | Правовий | Прийняття GDPR, введення адміністративної або кримінальної відповідальності за доступ до ресурсів, що блокуються на державному рівні | Реєстрація компанії в іншій державі у разі введення правової відповідальності. Переорієнтація на користувачів, які можуть не дотримуватись GDPR або дотримуватись частково. Забезпечення механізму дотримання GDPR у разі співробітництва з замовниками з ЄС. |
| 3 | Технічний | Зміна / оновлення / закриття / монетизація API соціальними мережами | Адаптація логіки системи, прийняття ризиків, у разі закриття API – міграція логіки на іншу соц. мережу. В якості превентивної міри може бути адаптація логіки таким чином, щоб якомога більше даних були змігровані у власне сховище. |

Результати аналізу факторів, що сприяють ринковому впровадженню проекту, повністю витікають з таблиці 5.4, а саме з колонки “Потреба, що формує ринок”. Основною реакцією компанії на дані потреби є їх задоволення з метою отримання прибутку для подальшого розвитку.

В таблиці 5.6 визначені загальні риси конкуренції на ринку.

Таблиця 5.6 – Результати аналізу загальних рис конкуренції на ринку

| Особливості конкурентного середовища | В чому проявляється дана характеристика |
|---|---|
| Вказати тип конкуренції – олігополія | Є декілька головних представників на ринку, що надають послуги, що базуються на аналізі користувачів за даними соц. мереж |
| 2. За рівнем конкурентної боротьби – світовий | На локальному ринку України відсутні компанії, що надають такий спектр послуг. На світовому ринку вони є: власники соціальних мереж, власники пошуковиків, а також компанії з більш спеціалізованим спектром аналізу користувачів |
| 3. За галузевою ознакою – внутрішньогалузева | Вузькоспеціалізований ринок конкретного ринку товару |
| 4. Конкуренція за видами товарів – товарно-родова, товарно-родова, товарно-видова | <p>Товарно-родова, тому що система може бути використана як додатковий модуль до існуючих систем таргетингу реклами. Також вона є платформою для виконання аналізу різного роду іншими учасниками ринку.</p> <p>Товарно-видова, тому що система може бути використана для аналізу користувачів з метою таргетингу реклами, що є прямою конкуренцією існуючих систем. Іншим варіантом використання є застосування її в рамках правоохоронної діяльності, що також є прямою конкуренцією існуючих аналогів.</p> |
| 5. За характером конкурентних переваг – нецінова | Основою є саме впровадження ноу-хау та нових засобів аналізу, які не надаються конкурентами, а не цінова політика за надані послуги |

Продовження таблиці 5.6

| | |
|---|--|
| 6. За інтенсивністю – марочна, не марочна | Основні конкуренти – компанії зі світовим ім'ям, які є олігополістами на ринку таргетованої реклами, тому неможливо не зтикнулися з впізнаваністю їх імен, отже боротьба з маркою також має місце. |
|---|--|

Результати аналізу умов конкуренції а галузі за моделлю п'яти сил М. Портера наведено в таблиці 5.7.

Таблиця 5.7 – Результати аналізу конкуренції в галузі за М. Портером

| Складові аналізу | Прямі конкуренти | Потенційні конкуренти | Постачальники | Клієнти | Товари-замінники |
|------------------|---|--|---|---|--|
| | Facebook Pixel, Google Adwords, Criteo, AgiOne і т.д. | БД Миротворець, Identigraf, InformNapalm | Facebook, Vk, Instagram, Google Cloud, AWS | Представники роздрібної торгівлі, правоохоронні органи, соціологічні організації | Реалізація подібної функціональності з боку компаній-власників соціальних мереж |
| Висновки: | Інтенсивність боротьби - критична | Можливості виходу на ринок присутні через надання унікального спектру послуг | Постачальники здатні диктувати деякі умови на ринку через надання API для отримання даних з соціальних мереж, а також через зміну спектру послуг, які надають їх інфраструктура | Клієнти можуть створити попит на специфічні дані, які можна отримати із соціальних мереж, а також на специфічні засоби аналізу даних. | Критичні обмеження у випадку оновлення пакету послуг іншими системами для таргетованої реклами, відсутні обмеження в інших напрямках |

На основі результатів аналізу конкуренції (таблиця 5.7), а також із урахуванням характеристик ідеї проекту (таблиця 5.1), вимог споживачів до товару (таблиця 5.5) та

факторів маркетингового середовища визначено перелік факторів конкурентоспроможності в таблиці 5.8.

Таблиця 5.8 – Обґрунтування факторів конкурентоспроможності

| № п/п | Фактор конкурентоспроможності | Обґрунтування (наведення чинників, що роблять фактор для порівняння конкурентних проектів значущим) |
|-------|---|--|
| 1 | Надання API для доступу до даних соц. мереж | Платформа інкапсулює логіку організації даних, уніфікує їх формат та надає єдине API для роботи з ними. При цьому дані можуть бути підготовлені із декількох соц. мереж |
| 2 | Створення механізму для аналізу даних | Замість надання кінцевого пакету послуг надається інструмент, який можна використовувати для виконання аналізу на основі даних соц. мереж. Цей механізм спрощує процес аналізу шляхом введення API для інтеграції з конвеєром обробки даних і автоматизуючи процес обробки користувачів на основі результатів класифікації окремо взятого профілю користувача і його публічних зв'язків. |

Результати SWOT-аналізу (матриці аналізу сильних та слабких сторін, загроз та можливостей) наведено в таблиці 5.9.

Таблиця 5.9 – SWOT-аналіз стартап-проекту

| | |
|--|--|
| Сильні сторони: - створення єдиної платформи та механізму для аналізу на основі даних соціальних мереж; | Слабкі сторони: - відсутність прямого доступу до даних; |
|--|--|

Продовження таблиці 5.9

| | |
|--|---|
| <ul style="list-style-type: none"> - автоматичний механізм ідентифікації осіб на зображеннях; - напівавтоматичний механізм класифікації осіб за зображеннями. | <ul style="list-style-type: none"> - залежність від запатентованих технологій; - відсутність значних обчислювальних ресурсів для розробки якісних класифікаторів та підтримки високої швидкодії; - відсутність інфраструктури для розвитку на ринкові таргетованої реклами. |
| <p>Можливості:</p> <ul style="list-style-type: none"> - Попит на нові рішення в сфері таргетування реклами; - відсутність єдиної платформи для виконання аналізу на основі даних соц. мереж; - суттєве збільшення попиту в Україні на системи пов'язані з безпекою інформаційного середовища. | <p>Загрози:</p> <ul style="list-style-type: none"> - Правова (прийняття GDPR, адміністративна/кримінальна відповідальність за доступ до заблокованих ресурсів); - політична (блокування ресурсів на державному рівні); - бізнес-аспект (зміни в політиці надання доступу до даних компаніями-власниками соціальних мереж); - дуже висока степінь конкуренції на ринку таргетованої реклами. |

Результати аналізу альтернатив ринкової поведінки наведено в таблиці 5.10.

Таблиця 5.10 – Результати аналізу альтернатив ринкової поведінки

| № п/п | Альтернатива (орієнтовний комплекс заходів) ринкової поведінки | Ймовірність отримання ресурсів | Строки реалізації |
|-------|--|--------------------------------|-------------------|
| 1 | Модель максимізації прибутку | Висока | Короткострокові |
| 2 | Модель максимізації обсягу продаж | Середня | Короткострокові |

Продовження таблиці 5.10

| | | | |
|---|--|--------|---------------|
| 3 | Модель забезпечення конкурентних переваг | Висока | Довгострокові |
| 4 | Модель максимізації доданої вартості | Висока | Довгострокові |

Вирішено обрати модель максимізації прибутку, тому що:

- висока імовірність отримати перші прибутки;
- строки реалізації є короткостроковими;
- кількість клієнтів може бути незначною, що дозволить визначити та проаналізувати недоліки системи, які не були знайдені на етапі підготовки до виходу на ринок;
- така поведінка дозволить дуже швидко отримати перші відгуки від клієнтів по роботі системи та процесам, що, в свою чергу, дозволить розробити подальшу стратегію розвитку проекту.

Як тільки проект буде стабілізований на ринку і будуть враховані всі перші відгуки по роботі системи, слід перейти до стратегії максимізації обсягу продаж, що дозволить стати бренду більш пізнаваним на ринку, отримати більше специфічних замовлень, на основі яких в майбутньому можна буде розробити стратегію конкурентних переваг, що дозволить увійти в сегменти, які вже зайняті конкурентами.

5.4 Розроблення ринкової стратегії проекту

В таблиці 5.11 наведені результати аналізу цільових груп потенційних споживачів.

Таблиця 5.11 – Вибір цільових груп потенційних користувачів

| № п/п | Опис профілю цільової групи потенційних клієнтів | Готовність споживачів сприйняти продукт | Орієнтовний попит в межах цільової групи (сегменту) | Інтенсивність конкуренції в сегменті | Простота входу у сегмент |
|-------|--|---|---|--------------------------------------|--------------------------|
| | | | | | |

Продовження таблиці 5.11

| | | | | | |
|---|---|------------------------|--------------------|-------------|--------------------|
| 1 | Наукові організації, держстат та інші статистичні організації, соціологічні організації | Висока готовність | Високий попит | Відсутня | Середня складність |
| 2 | Представники ринку електронної торгівлі | Середня готовність | Високий попит | Дуже висока | Дуже складно |
| 3 | Правоохоронні органи | Дуже висока готовність | Дуже високий попит | Відсутня | Середня складність |
| Які цільові групи обрано: вирішено в першу чергу орієнтуватись на правоохоронні органи, а також будь-які організації, які зацікавлені в існуванні платформи для аналізу даних користувачів соціальних мереж (соціологічні та наукові організації) | | | | | |

В таблиці 5.12 наведено результати визначення базової стратегії розвитку проекту.

Таблиця 5.12 – Результати визначення базової стратегії розвитку проекту

| № п/п | Обрана альтернатива розвитку проекту | Стратегія охоплення ринку | Базова стратегія розвитку* |
|-------|--|--|---------------------------------|
| 1 | Модель максимізації прибутку | Стратегія диференційованого маркетингу | Стратегія спеціалізації |
| 2 | Модель максимізації обсягу продаж | Стратегія диференційованого маркетингу | Стратегія спеціалізації |
| 3 | Модель забезпечення конкурентних переваг | Стратегія масового маркетингу | Стратегія диференціації |
| 4 | Модель максимізації доданої вартості | Стратегія масового маркетингу | Стратегія лідерства по витратах |

В таблиці 5.13 наведено результати вибору стратегії конкурентної поведінки.

Таблиця 5.13 – Результати вибору стратегії конкурентної поведінки

| № п/п | Чи є проект «першопрохідцем» на ринку? | Чи буде компанія шукати нових споживачів, або забирати існуючих у конкурентів? | Чи буде компанія копіювати основні характеристики товару конкурента, і які? | Стратегія конкурентної поведінки |
|-------|--|--|---|--------------------------------------|
| 1 | Так | Нові користувачі | Ні | Стратегія зайняття конкурентної ніші |

Висновки до розділу

Результати аналізу різних аспектів розроблення стартап-проекту:

- наявний попит та динаміка ринку є позитивною;
- проект є технологічно здійсненним;
- є перспективи впровадження, адже присутні сегменти з низькою конкуренцією та бар'єром входження;
- є механізми для боротьби з ризиками;
- проект є конкурентоспроможним, адже надаватиме послуги, яких не існує на ринку;
- в якості базової стратегії розвитку вирішено обрати стратегію спеціалізації у короткостроковій перспективі та стратегії диференціації та лідерства по витратах у довгостроковій перспективі;
- в якості стратегії конкурентної поведінки обрана стратегія зайняття конкурентної ніші.

ЗАГАЛЬНІ ВИСНОВКИ

Основні результати роботи полягають в наступному:

- а) запропоновано і реалізовано модифікацію методу порівняння зображень, результатом роботи якої є процентне співвідношення схожості зображень. Результати порівняння зображень даним методом є більш правильні, аніж у методу, на якому базується модифікація;
- б) запропоновано і реалізовано метод групування зображень, який автоматично визначає кількість груп;
- в) запропоновано і реалізовано алгоритм класифікації користувачів за їх зв'язками, що базуються на спільній присутності на зображеннях;
- г) реалізована система класифікації осіб за даними соціальних мереж, архітектура якої дозволяє незалежно масштабувати кожен її компонент, а окремі компоненти були реалізовані з використанням технологій обробки надвеликих масивів даних.

Подальші напрямки досліджень мають бути направлені на розвиток методів і технологій, які були використані в магістерській дисертації, а саме:

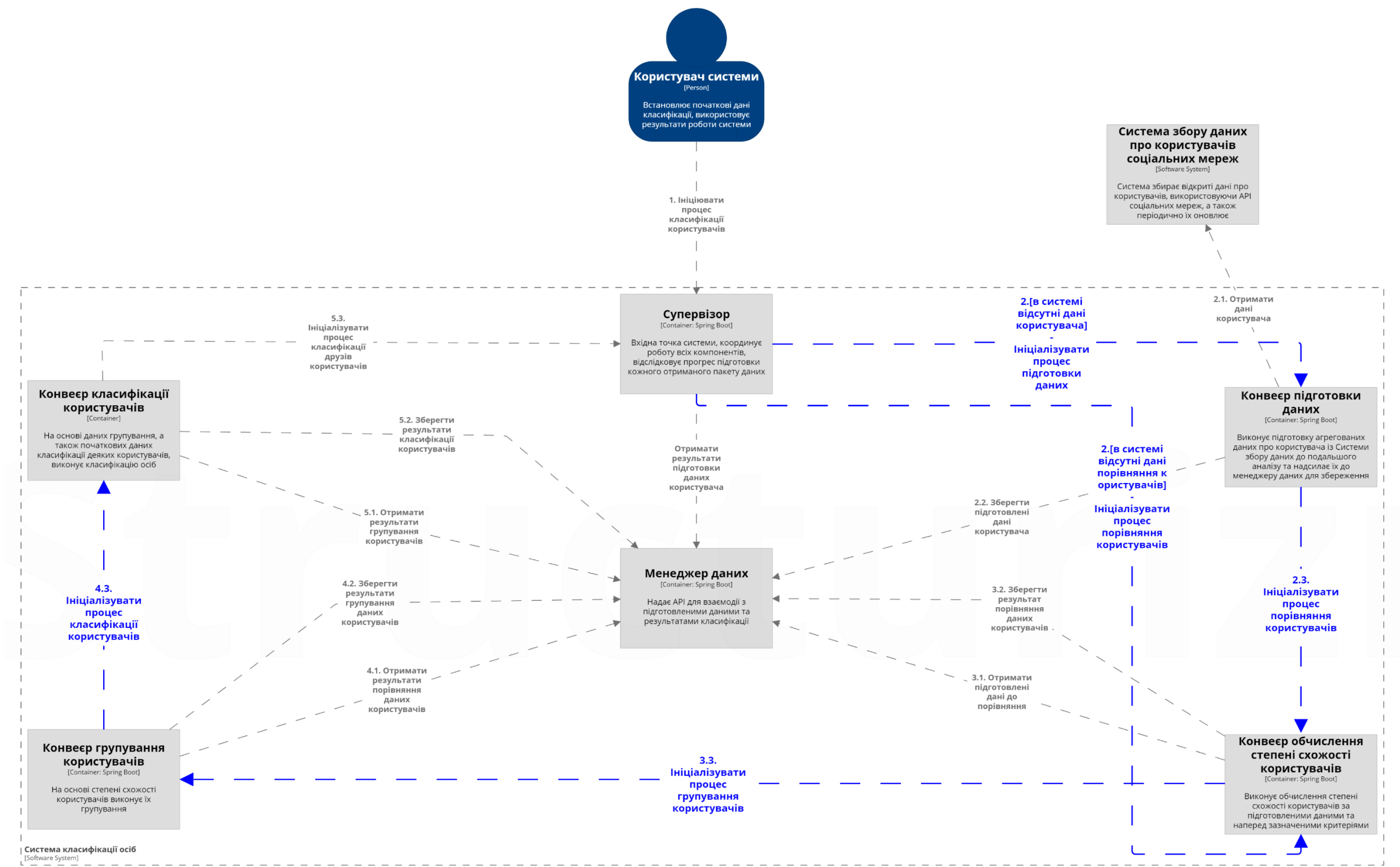
- а) критеріїв подібності зображень, адже від якості їх роботи залежать точності групування та класифікації, як показали результати експериментів;
- б) методів групування даних, в першу чергу досліджувати способи зменшення їх асимптотичної складності, засобів побудови підграфів, що не мають спільних вершин, з повних підграфів;
- в) модифікації запропонованого алгоритму класифікації таким чином, щоб результатом роботи був процент приналежності користувача соціальних мереж до кожного класу. Це дозволить мати більш чітке уявлення про структуру спільнот користувачів за даними їх спільної присутності на зображеннях, а також встановлення пріоритетів їх обробки.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

- 1) Grayscale to RGB Conversion [Електронний ресурс] – Режим доступу до ресурсу: https://www.tutorialspoint.com/dip/grayscale_to_rgb_conversion.htm
- 2) Concept of Blurring [Електронний ресурс] – Режим доступу до ресурсу: https://www.tutorialspoint.com/dip/concept_of_blurring.htm.
- 3) Гонсалес Р. С. Цифровая обработка изображений / Р. С. Гонсалес, Р. Е. Вудс. – Москва: Техносфера, 2012.
- 4) Local Feature Detectors and Descriptors [Електронний ресурс] – Режим доступу до ресурсу: http://www.cse.iitm.ac.in/~vplab/courses/CV_DIP/PDF/Feature_Detectors_and_Descriptors.pdf
- 5) Distinctive Image Features from Scale-Invariant Keypoints [Електронний ресурс] / David G. Lowe – Режим доступу до ресурсу: <https://www.cs.ubc.ca/~lowe/papers/ijcv04.pdf>
- 6) Face recognition using Gabor waveleth transform [Електронний ресурс] / Burcu Keremekci – Режим доступу до ресурсу: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.100.6259&rep=rep1&type=pdf>
- 7) Novel Matching Methods for Automatic Face Recognition using SIFT [Електронний ресурс] / Ladislav Lenc, Pavel Král – Режим доступу до ресурсу: https://home.zcu.cz/~pkral/papers/kral_aiai12.pdf
- 8) Automatic face recognition system based on the SIFT features [Електронний ресурс] / Ladislav Lenc, Pavel Král – Режим доступу до ресурсу: https://home.zcu.cz/~pkral/papers/cee_kral15.pdf
- 9) Person-Specific SIFT Features for Face Recognition [Електронний ресурс] / Jun Luo, Yong Ma, Erina Takikawa, Shihong Lao, Masato Kawade, Bao-Liang Lu – Режим доступу до ресурсу: <https://ieeexplore.ieee.org/document/4217478>

- 10) Bron, C., & Kerbosch, J. Algorithm 457: finding all cliques of an undirected graph. – Communications of the ACM. 1973. – 575–577.
- 11) The worst-case time complexity for generating all maximal cliques and computational experiments [Электронный ресурс] / Tomita, E., Tanaka, A., & Takahashi, H. – Режим доступа до ресурсу: <https://snap.stanford.edu/class/cs224w-readings/tomita06cliques.pdf>
- 12) A note on the problem of reporting maximal cliques [Электронный ресурс] / Cazals, F.; Karande, C. – Режим доступа до ресурсу: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.833.8192&rep=rep1&type=pdf>
- 13) Listing All Maximal Cliques in Sparse Graphs in Near-optimal Time [Электронный ресурс] / David Eppstein, Maarten Löffler, and Darren Strash – Режим доступа до ресурсу: <https://arxiv.org/pdf/1006.5440.pdf>
- 14) Near linear time algorithm to detect community structures in large-scale networks [Электронный ресурс] / Usha Nandini Raghavan, R'eka Albert and Soundar Kumara – Режим доступа до ресурсу: <https://arxiv.org/pdf/0709.2938.pdf>

ДОДАТОК А ФУНКЦІОНАЛЬНА СХЕМА ПРОЦЕСУ КЛАСИФІКАЦІЇ КОРИСТУВАЧА



Демонстраційний плакат до магістерської дисертації

Функціональна схема процесу класифікації користувача

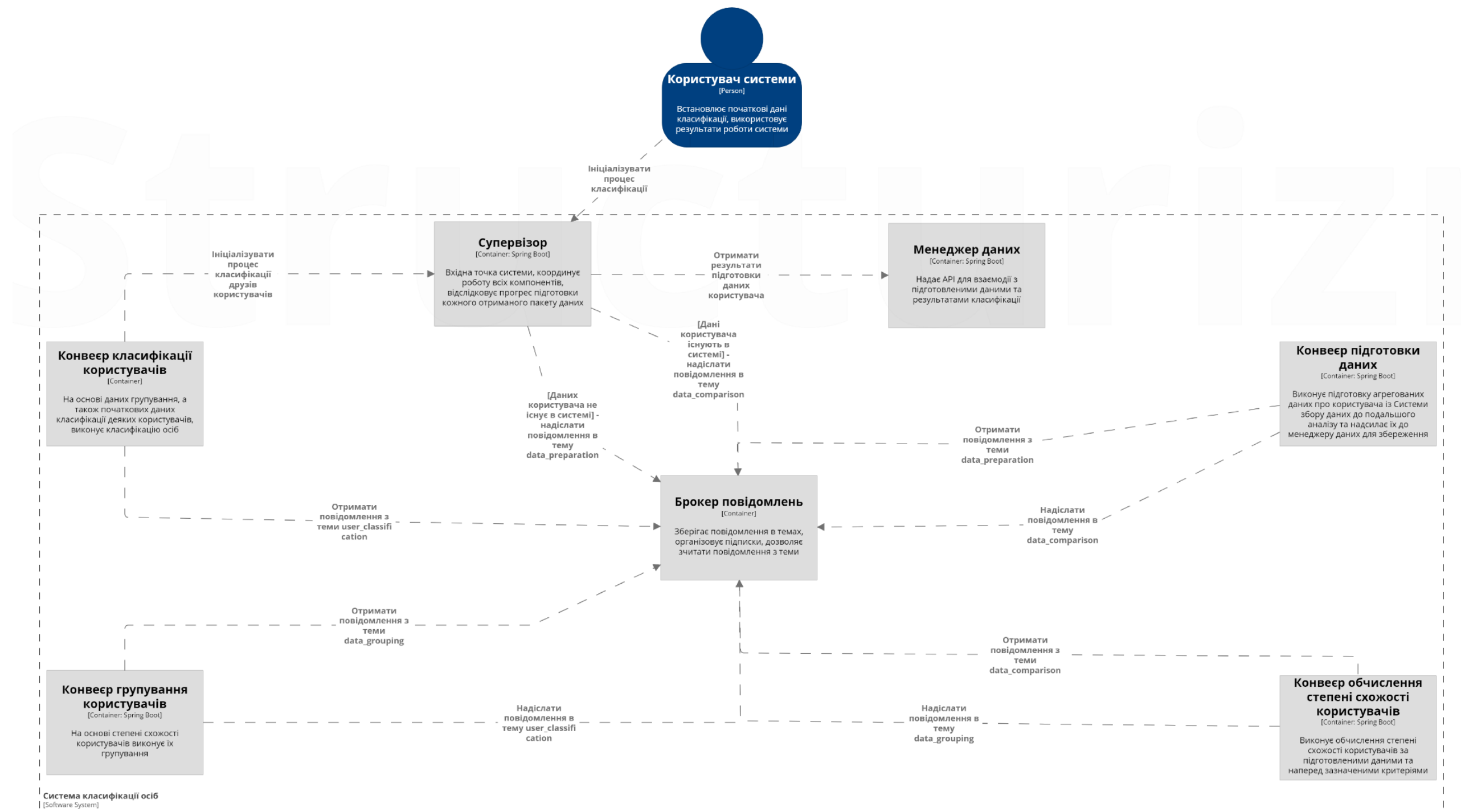
Виконав студент гр. ІІІ-82мпв

Кахерський О.І.

Керівник

Гавриленко О.В.

ДОДАТОК Б ФУНКЦІОНАЛЬНА СХЕМА ПРОЦЕСУ ПЕРЕДАЧІ УПРАВЛІННЯ ПІД ЧАС КЛАСИФІКАЦІЇ КОРИСТУВАЧА



Демонстраційний плакат до магістерської дисертації

Функціональна схема процесу передачі управління під час класифікації користувача

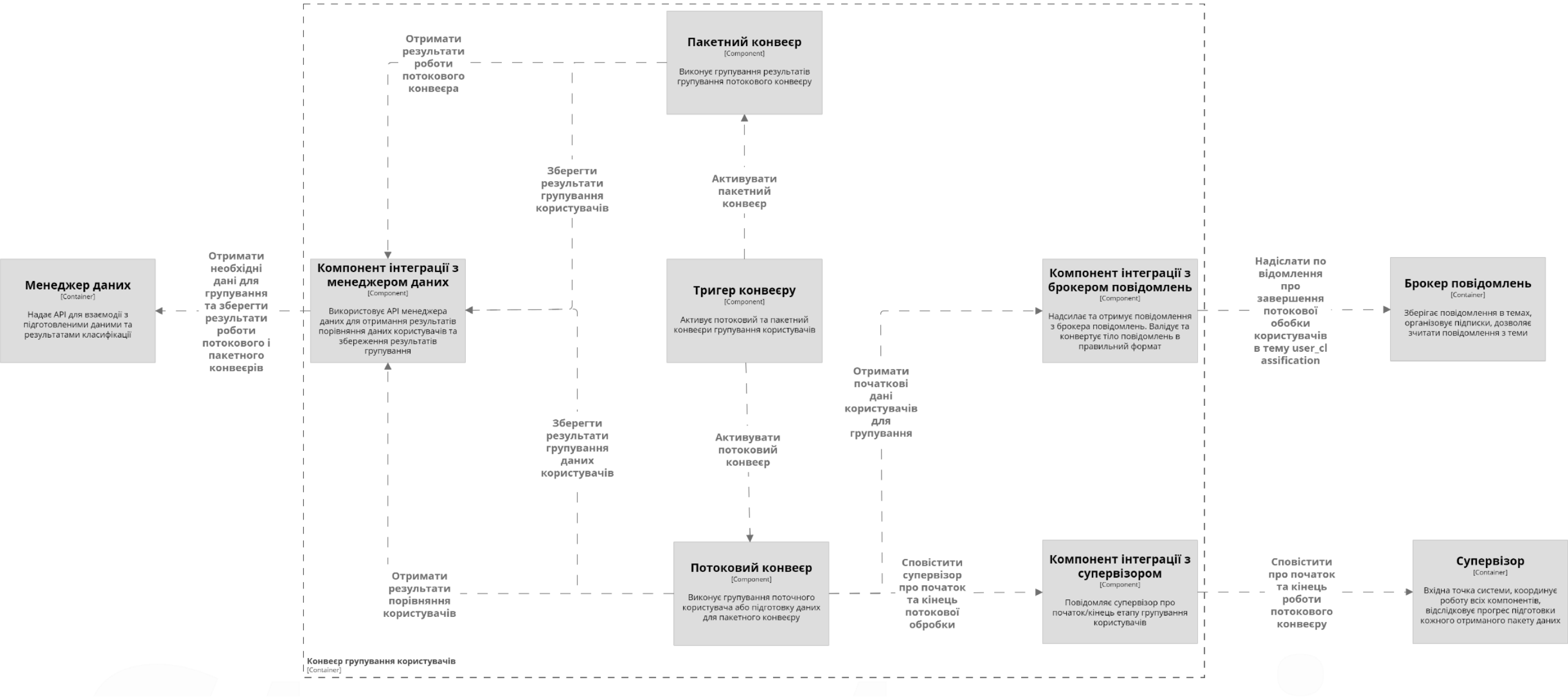
Виконав студент гр. ІІІ-82мпв

Кахерський О.І.

Керівник

Гавриленко О.В.

ДОДАТОК В СТРУКТУРНА СХЕМА СЕРВІСУ ГРУПУВАННЯ КОРИСТУВАЧІВ



Демонстраційний плакат до магістерської дисертації

Структурна схема сервісу групування користувачів

Виконав студент гр. ІІІ-82мпв

Кахерський О.І.

Керівник

Гавриленко О.В.